人工智能全传

让牛津大学计算机学院院长告诉你
人工智能会不会超越人类



[英] 迈克尔·伍尔德里奇 著
（牛津大学计算机学院院长）
许舒 译

THE ROAD TO CONSCIOUS MACHINES
THE STORY OF ARTIFICIAL INTELLIGENCE

浙江科学技术出版社

VISIT...

LANZAROTE Caliente.COM

人工智能全传

让牛津大学计算机学院院长告诉你
人工智能会不会超越人类



[英] 迈克尔·伍尔德里奇 著
（牛津大学计算机学院院长）
许舒 译

THE ROAD TO CONSCIOUS MACHINES
THE STORY OF ARTIFICIAL INTELLIGENCE

浙江科学技术出版社

人工智能全传

[英] 迈克尔·伍尔德里奇　著
许　舒　译

浙江科学技术出版社

著作合同登记号 图字：11-2020-459 号

Copyright © 2020 by Michael Wooldridge
This edition arranged with Felicity Bryan Associates Ltd. through Andrew Nurnberg Associates International Limited

中文版权：©2021 读客文化股份有限公司
经授权，读客文化股份有限公司拥有本书的中文（简体）版权

图书在版编目（CIP）数据

人工智能全传 /（英）迈克尔·伍尔德里奇（Michael Wooldridge）著；许舒译．—杭州：浙江科学技术出版社，2021.2

书名原文：The Road to Conscious Machines: The Story of AI

ISBN 978-7-5341-9455-9

Ⅰ．①人… Ⅱ．①迈… ②许… Ⅲ．①人工智能－普及读物 Ⅳ．① TP18-49

中国版本图书馆 CIP 数据核字 (2021) 第 016261 号

书　　名　人工智能全传
著　　者　［英］迈克尔·伍尔德里奇
译　　者　许　舒

出　　版	浙江科学技术出版社	网　　址	www.zkpress.com
地　　址	杭州市体育场路 347 号	联系电话	010-87681002
邮政编码	310006	印　　刷	天津联城印刷有限公司
发　　行	读客文化股份有限公司		

开　　本	880 × 1230　1/32	印　　张	12
字　　数	234 000		
版　　次	2021 年 2 月第 1 版	印　　次	2021 年 2 月第 1 次印刷
书　　号	ISBN 978-7-5341-9455-9	定　　价	78.00 元

特邀编辑　肖　飒　　刘昀琪　　张伊珂
责任编辑　卢晓梅　　责任校对　张　宁　　责任美编　金　晖　　责任印务　叶文炀
封面设计　汪芝灵

图书出现倒装、缺页等印装质量问题，本社销售部负责调换

本书献给牛津大学赫特福德学院的院长、研究员及学者们

目录

第一部分 人工智能是什么
第一章 图灵的电子大脑
第二部分 我们是怎么走到这一步的
第二章 黄金年代
第三章 知识就是力量
第四章 机器人与其合理性
第五章 深度突破
第三部分 我们将去向何处
第六章 人工智能的今天
第七章 杞人忧天——我们想象中的人工智能会出什么错
第八章 现实中的人工智能会导致什么问题
第九章 通往有意识的机器之路
后记
致谢
词汇表
附录
附录A 理解规则
附录B 理解PROLOG
附录C 理解贝叶斯定理
附录D 理解神经网络
延伸阅读
作者注

第一部分 人工智能是什么

第一章 图灵的电子大脑

我建议大家考虑一个问题：“机器能思考吗？”

——艾伦·图灵，1950年

每个故事都有个开头，对人工智能(Artificial Intelligence，AI)而言，可以选择的开头有很多，因为人类对人工智能的梦想，可以追溯到很久很久以前。

我们可以选择从古希腊开始，铁匠之神赫菲斯托斯(Hephaestus)拥有赋予金属物品生命的能力。

我们可以选择从16世纪的布拉格开始，传说中伟大的拉比在那里用黏土制作了一个傀儡魔像，意图保护该城市的犹太人免遭迫害。

我们可以选择从18世纪苏格兰的詹姆斯·瓦特(James Watt)开始，他为正在建造的蒸汽机设计了一个巧妙的自动控制系统——调速器，从而为现代控制理论奠定了基础。

我们可以选择从19世纪初开始，年轻的玛丽·雪莱(Mary Shelley)被恶劣的天气困在瑞士的一栋别墅里，和她的丈夫——诗人珀西·比希·雪莱(Percy Bysshe Shelley)，以及他们家的朋友拜伦(Byron)勋爵一起，玩了一场文字游戏，创作了《弗兰肯斯坦》[1]。

我们可以选择从19世纪30年代开始，在彼时的伦敦，阿达·洛芙莱斯(Ada Lovelace)，上文提到的那位拜伦勋爵的女儿——虽然父女俩关系非常疏远——与查尔斯·巴贝奇(Charles Babbage)建立了友谊，性情乖张的巴贝奇发明了分析机，这激励着才华横溢的阿达探寻机器是否最终能够具备创造性[2]。

我们也可以选择从18世纪掀起的自动机器装置狂热开始。那些小型自动机器装置被设计得无比精妙，有一种被赋予生命的错觉。

追溯人工智能的开端，我们可以有许许多多的选择，但对我来说，人工智能故事的开头与计算机的开端应该是一致的。对此，我们有一个非常明确的起点：1935年，剑桥大学国王学院，一位才华横溢、不拘泥于传统的年轻学生——艾伦·图灵(Alan Turing)。

剑桥1935

现在，艾伦·图灵是一个家喻户晓的名字，但我们很难想象在20世纪80年代，这个名字在数学和计算机领域以外压根不为人知。虽然相关专业的学生可以在教材上偶遇图灵的名字，但对他非凡的成就知之不详，更不了解他那悲惨的命运。部分原因是图灵在第二次世界大战期间为英国政府效力，承担了一项非常重要的工作，而这项杰出工作的成果一直秘而不宣，直到20世纪70年代才公开[1]。但是，更重要的原因是歧视，因为图灵是同性恋，在当时的英国，同性恋是一种犯罪。1952年，他以“严重猥亵罪”被起诉和定罪，被迫服用某种旨在降低性欲的粗制滥造的药物来进行所谓的“化学阉割”。两年后，他亲手结束了自己的生命，年仅41岁[2]。

尽管如今的我们对图灵的故事所知甚少，但多少略知一二。其中最著名的桥段自然是二战期间他在布莱切利庄园破译密码的传奇，因2014年好莱坞电影《模仿游戏》而天下皆知(虽然电影和现实的差距大得惊人)。他为盟军的最终胜利立下了汗马功劳。但是，人工智能研究人员和计算机科学家崇拜图灵，是出于完全不同的理由：他是真正意义上的计算机发明者，不久，又成为人工智能领域的主要奠基人。

图灵在诸多领域都有非凡成就，最非凡的是在偶然的机会下，他发明了计算机。20世纪30年代中期，图灵还是剑桥大学数学系的一名学生，他给自己定下一个超前的挑战，那就是解决彼时最重要的终极数学问题之一：判定问题。这个令人印象深刻——坦率地说，是可怕——的问题，由数学家大卫·希尔伯特(David Hilbert)于1928年提出。判定问题是指，是否所有的数学描述都是“可判定的”，即是否存在一种方法，可以确认某个给定的数学描述是真是假。当然，希尔伯特关心的领域并非“有没有上帝”或者“生命的意义是什么”之类，而是数学领域中的判定问题。判定问题是一种有着明确“是”或“否”的答案的数学问题。以下是判定问题的示例：

·2 + 2是否等于4？

·4×4是否等于16？

·7919是否是一个质数？

很凑巧，这几个问题的答案都是“是”。前两个问题的答案是显而易见的，但是，除非你是个痴迷于质数的人，否则要回答第三个问题得稍微费点儿劲了。所以，我们来看看最后一个问题。

你应该记得，质数的定义是一个只能被自己和1整除的整数。现在回到这个问题，我相信从原则上讲，你能够寻找到一个验证它的方法。有一个很容易想到的笨办法，当然，对7919这么大的数字来说有些烦琐：依次用每一个整数去除它，看看它能不能被某个数除尽。重复这个过程以后，你会发现，除了1和7919本身，没有其他可以除尽它的数，这就意味着7919确实是一个质数[3]。

重点在于，上述问题可以用一个精准的方式来寻找答案。这种方式不需要任何灵活的应用——只是一些需要死记硬背的方法。要寻找答案，我们只需要严格按照方法所示的步骤执行就可以。

既然有这样的方法保证能够验证这个问题的是与否(只要有足够的时间)，我们就可以说，“某个数是否是一个质数”是可判定问题。在此我强调一下，这就意味着，每当我们遇见“数n是否是一个质数”这样的问题时，只要给予足够的时间，就能找到答案：遵循相关步骤执行，最终就能得到明确的答案。

现在，判定问题提出：是否所有的数学判定问题都是可判定的？或者说，是否存在某种数学问题，无论你投入多少时间，都无法寻找到相应的可供严格遵循来解决问题的方法？

这是一个基础法则类问题，本质是数学问题是否都可以简化为寻找对应算法就可以解决的问题。回答这个基础问题是图灵在1935年给自己立下的艰巨挑战，而他以令人难以置信的惊人速度解答出来了。

一想到深奥的数学问题，我们很容易就会认为解决它们的方法是冗长而复杂的，涉及大篇幅的方程式和证明过程。有时候，事实的确是这样，例如英国数学家安德鲁·威尔斯(Andrew Wiles)在20世纪90年代论证著名的费马大定理，数学界花费了好几年时间才消化了他那几百页的手稿，并确信他的论证是正确的。按照这样的标准，图灵解决判定问题的

方案简直是颠覆性的。

与大家想象中的不同，图灵的证明短小精悍，而且很容易读懂(在确立了基本框架以后，真正的证明实际上只有几行)。重点在于，图灵意识到，他需要一个精准定义解决问题方法的方案，为此，他发明了一种可以解决数学问题的机器，如今，为了纪念他，我们称之为图灵机。图灵机是一种对方法的具体描述，比如上文提到的检查质数的步骤。图灵机的作用是严格遵循既定的步骤执行运算。在此，我要强调的是，尽管图灵将它称为“机器”，但在那时它只是个抽象的数学概念。通过发明一台机器来解决一个艰深的数学问题，这种想法相当颠覆传统，当年的数学家们肯定都被搞迷糊了。

图灵机就如强悍的神兽，任何你能想出来的数学方法都可以被编码成图灵机。因此，如果所有的判定问题都是可以解决的，就意味着任何判定问题都可以通过设计一个专用的图灵机来解决。也就是说，为了解答希尔伯特的问题，你所要做的是，证明存在某种判定问题是任何图灵机都无法解决的。这正是图灵证明这一命题的方法。

接下来，图灵耍了点小把戏，让他的机器摇身一变，成为能够解决通用问题的机器。他设计了一种图灵机，可以按照人们给它的任意方法进行运算，现在我们把它称为通用图灵机[4]。一台计算机最核心也最基础的部分，就是一台真正的通用图灵机。计算机所运行的程序，本质上就是运算的步骤，类似前文我们提到的确认质数的步骤。

尽管这不是我们故事的重点，但至少得提出来图灵是怎么用它的新发明解决判定问题的。除了惊叹于他精巧独到的心思，我们还应该意识到，它跟人工智能最终是否可能存在有着密切关系。

图灵的构想是存在这样一台图灵机，可以用来判定任意图灵机的运行结果。他考虑了以下问题：给定一台图灵机和相关输入集，它最终是会停止，还是永无止境地运行下去？这就是一个判定问题了，属于我们前文讨论过的范畴，尽管它相对复杂一些。现在，假设存在一个机器能够判定这个判定问题，图灵指出，这个假设会引起悖论[3]。因此，没有办法检测出图灵机是否停止。那么，“图灵机是否停止”是一个不可判定问题。所以图灵得出结论：存在某些判定问题不能简单地按照确定的步骤来解决。他解决了希尔伯特的难题：数学并不能被简化为遵循方法解决问题[5]。

这一结果是20世纪数学界最伟大的成就之一，单凭它就足以让图灵在数学界名垂青史。但更伟大的是它的副产物——通用问题解决机器，即图灵机。图灵发明图灵机的时候，它只是个抽象的概念，他并没有想着将它实体化。不过没过多久，不少人，包括图灵自己，开始着手把这个想法转化成现实。在二战时的慕尼黑，康拉德·楚泽(Konrad Zuse)为德国航天部设计了一台名为Z3的计算机，虽然它算不上一台完整的计算机，但引入了不少关键部分。在大西洋彼岸的美国宾夕法尼亚州，由约翰·穆克里(John Mouchly)和普雷斯伯·埃克特(J. Presper Eckert)领导的小组开发了一台名为ENIAC的机器来计算火炮射击表。杰出的匈牙利裔数学家约翰·冯·诺依曼(John von Neuman)对它进行了相关调整，使ENIAC具备现代计算机的基本架构(为了纪念这位数学家，传统计算机的架构被称为“冯·诺依曼架构”)。在战后的英国，弗雷德·威廉姆斯(Fred Williams)和汤姆·基尔伯恩(Tom Kilburn)建造了昵称为“曼彻斯特宝贝”的小规模实验机，直接促成了世界上第一台商用计算机“费兰蒂一号”的出现(图灵本人于1948年加入曼彻斯特大学的工作团队，并编写了最早运行的程序)。

到了20世纪50年代，现代计算机的所有关键部件都被发明出来了。图灵机已经从数学概念转化为实实在在的机器——只要你有足够的金钱买下它，以及足够的场地摆放它(要装下“费兰蒂一号”，至少需要两个16英尺[4]长、8英尺高、4英尺宽的储藏室。机器的功率为27千瓦——它消耗的电能可以供应至少三个现代家庭用电)。当然，随着时代的发展，计算机变得越来越小巧，越来越便宜。

电子大脑的实际功能

还有什么比耸人听闻的标题更能取悦报社编辑的呢？第二次世界大战结束后，第一台计算机建成，世界各地的报纸都争相报道这项神奇的发明——电子大脑。这些看起来可怕又复杂的机器有着令人炫目的计算能力，用它们处理海量的复杂算术问题，其速度和精确程度远远超乎人类想象。对于那些不了解计算机原理的人来说，能完成复杂算术任务的机器似乎拥有某种高级智能。因此，它被人们称为电子大脑，然而，名不副实(直到20世纪80年代，当我首次对计算机领域产生兴趣时还能听到这类说法)。事实上，这些电子大脑确实能承担许多对人类而言烦琐、复杂且困难的计算，但它们并没有智能可言。所以，了解清楚计算机最初发明出来是做什么的，以及它不能做什么，是理解人工智能局限的核

心，也能让我们明白，为什么实现人工智能的宏伟目标如此困难。

请记住，图灵机，以及它的物理表现形式计算机，它们只是遵循各种指令的机器而已。这是它们存在的唯一目的——它们被设计出来就是做这个的，也只会做这个。我们给予图灵机的指令，现在被称为算法或者程序[6]。大多数程序员可能都不太清楚他们打交道的东西本质上跟图灵机类似，这也不怪他们——直接为图灵机编程简直是受罪，无聊、烦躁得令人生厌，一代又一代在编程路上吃尽苦头的计算机专业学生能够证明这一点。因此，我们在图灵机上构建更高级的语言，诸如Python、Java和C语言之类，让编程变得简单。高级语言的作用是向程序员隐藏机器语言中某些烦琐到可怕的细节，让编程变得容易一些。但从本质上来说，编程仍然是枯燥乏味、令人生厌的，这就是为什么学编程这么难，为什么计算机程序总是莫名其妙地崩溃，为什么优秀的程序员薪水总是这么优渥。

在本书中我不会教你编程，不过了解一下程序指令的作用以及电脑怎么按照步骤来执行程序也是有必要的。粗略地说，计算机能做的，就是按步骤执行一系列指令而已[7]，例如：

将A与B相加

如果结果大于C，则执行D，否则执行E

重复执行指令F，直到遇见情况G

所有的计算机程序都能归结为类似的指令列表，不管是Microsoft Word还是PowerPoint，不管是《使命召唤》还是《我的世界》，不管是脸书、谷歌还是淘宝，不管是浏览器还是手机App，抑或是支付宝、微信、QQ……全都能归结为类似的指令列表。如果我们要制造智能机器，它的智能最终必须缩减到遵从这些简单的、明确的指令上。这从本质上对人工智能提出了挑战：把这些简单的指令排列组合起来，真的可以产生智能行为吗？

在本章的剩余部分，我会深入挖掘这个问题，并试图明确它在人工智能发展历程上所产生的影响。然而，在讲述之前，我得先为计算机正名。本章截至目前，我向你描述的计算机似乎就是一坨无用的废铁，我觉得有必要强调一下它的伟大之处，以免误导大家。

首先，计算机的运算速度非常快，非常、非常、非常快。虽然这点

尽人皆知，但是我们在日常生活中很难直观地感受到。所以，我们来量化一下这个陈述。在我写这本书的时候，一台普通的台式机以全速运行，每秒可以处理1000亿条指令。1000亿大概是银河系所有恒星的数量，不过这么说还是不太直观。所以，请你想象一下自己要通过手动执行指令来和计算机进行一次较量。你大概每10秒执行一条指令。你得做到不吃不喝、不眠不休(全年365天，全天24小时，每小时60分钟，每分钟60秒)，那么大概需要31 710年，你才能完成计算机1秒钟就能搞定的工作。

当然，除了速度慢得令人发指以外，跟计算机比起来，你还有一个最关键的劣势：你不可能在执行海量任务的时候保证不出错。相比人类，计算机的错误率极低。当然，程序崩溃是常见的，但那几乎全是程序员编写程序时出的错，而不是计算机本身的问题。现代计算机的处理器可靠性非常高，它们的平均无故障运行时间高达50 000小时，每秒钟都能忠实地执行数百亿条指令。

最后，虽然计算机只是遵循指令的机器，但并不意味着它不能做决定。计算机当然可以做决策，只是我们必须给出它做决策所需要的精准指令。计算机随后可以自行调整这些指令，只要我们指导它在何种情况下应该如何做——这就意味着，计算机可以随着时间推移改变其行为——它能够学习。

人工智能的产生为何如此艰难

现在我们了解到，计算机可以非常迅捷而精准地执行简单的指令，另外，只要输入精准定义过的指令，它就能够做决策。我们可以顺理成章地得出结论，某些需要计算机为我们处理的事务可以用非常简单的方式进行编码，可有些功能则不然。要理解人工智能的产生为何如此艰难，为什么如此难以定义何为人工智能真正的“进步”，我们需要了解一下到底哪些问题是容易用编程来解决的，而哪些问题则很难用编程解决，以及为什么。图1展示了我们希望计算机能够完成的任务，以及它们的难易程度及实现时间。

排在最前面的是计算。让计算机做计算是最简单不过的，因为所有的基础运算(加、减、乘、除)都可以用非常简单的步骤来运行——你在学校里都学过，哪怕现在已经不记得了。这些步骤可以直接编译成计算机程序，早期的计算机本身就是为了解决计算问题而生的(1948年，图

灵加入曼彻斯特大学团队，为计算机“曼彻斯特宝贝”编写的第一条程序就是执行长除法。在解决了20世纪最深刻的数学问题之一以后，又回到了学校里学过的基础数学上来，这对图灵而言定然是一次奇特的经历)。

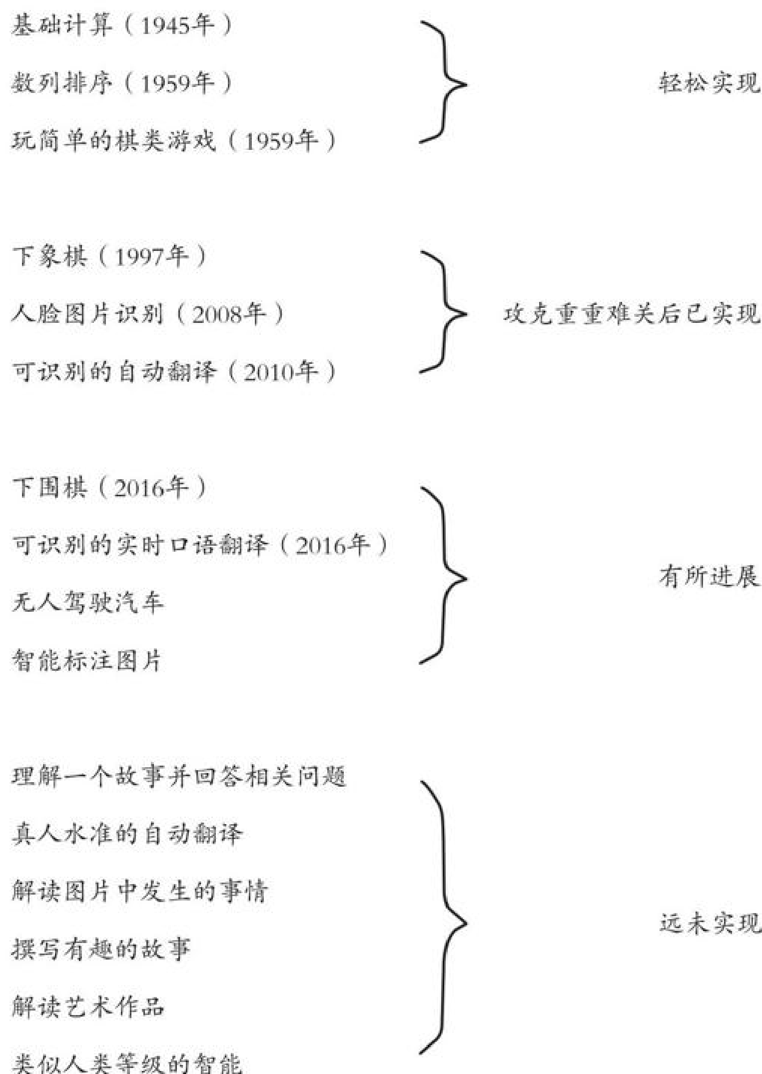


图1　我们希望计算机能完成的任务，按照难易程度排序括号中的年份表示问题解决的大致时间，而目前为止，对于“远未实现”的任务，我们毫无头绪。

接下来是排序，是指把一系列随机的数字按照升序排列，或者一系

列姓名按照字母顺序排列。这听起来一点儿都不人工智能，事实上排序的步骤并不复杂。然而，传统的笨办法慢得令人发指。一直到1959年，快速排序法才被发明出来，排序这个老大难问题才算得到了一个有效率的解决方法(快速排序法诞生后50年内，都没出现比它更优秀的排序算法)[8]。

然后我们要讨论需要攻克重重难关才能解决的问题。下棋就是重大挑战之一，它也是人工智能故事中非常重要的一环。事实上，有一种简单粗暴的方式可以玩好棋类游戏，基于一种名为搜索的技术，我们将在下一章详细讨论它。问题在于，尽管基于搜索进行编程是最简单不过的事情，但它除了能适应一些简单的游戏以外毫无用处，因为它需要占用太多内存，消耗太多时间——哪怕我们把整个宇宙的每一个原子都用来建造计算机，它也承担不起用“简单粗暴”的搜索方式来下一盘围棋或者象棋。为了让搜索具有可行性，我们需要增加一些额外的东西——就如我们在下一章即将看到的那样，至此，人工智能开始登场。

我们知道有办法解决问题，但是在实践技术中却无法实施——因为它们需要的计算资源太过庞大。这种情况在人工智能研究领域司空见惯，围绕着怎么处理这类问题，人们也进行了大量的研究。

不过列表中接下来的问题并非此类，人脸识别、自动翻译和实时可用的口语翻译与棋类游戏不一样，传统的计算机技术无法为我们提供解决这类问题的方法，我们需要探索全新的解决方案。当然，在列表上，这类问题已经通过一种名为机器学习的方式得到了解决，我们将在后续章节深入了解有关机器学习的内容。

接下来是一个吸引人的话题，无人驾驶汽车。驾驶对人类而言是一件很简单的事情，开车并不是一项需要高智商的技能。但事实证明，让电脑来控制汽车简直困难重重。主要的问题在于，汽车需要知道它所处的位置以及周围环境。想象一下，一辆无人驾驶的汽车来到纽约繁忙的十字路口，数不尽的汽车从它身边呼啸而过，还有行人、自行车、道路施工、交通标志和各种标线。还有可能遇到下雨、下雪或者大雾，这让周围情况变得更加复杂(在纽约，这三种气象甚至可能同时出现)。这种情况下，难点并不在于你要做什么(减速、加速、左转或者右转等)，而是要弄清你的周围发生了什么，即将发生什么——确认你的位置，周围有哪些车辆，它们的位置，它们会朝哪个方向行驶，行人的位置，行人的移动轨迹，等等。如果充分掌握了所有信息，那么你决定接下来该怎

么做就非常容易了。(我们会在稍后的章节里面讨论无人驾驶的细节。)

然后就是我们真不知道该怎么解决的问题了：计算机怎么去理解一个复杂的故事，并且回答有关它的问题？计算机如何能把类似小说这种细致入微的文本翻译到信达雅的程度？计算机如何看图说话——不光是识别图中的人物，而是要诠释图中发生的事情？计算机如何能创作出一个有趣的故事，或者解读一件艺术品，比如一幅画？最后，仍然回到我们的“宏伟梦想”上，计算机如何拥有类似人类的智能？

因此，人工智能的进步意味着让计算机能够越来越多地完成图1中所列举的任务，这些任务难以完成的原因通常有两种。第一种是我们虽然知道理论上有某种可以解决问题的方法，但实际上却行不通，因为它需要太过漫长的运算时间，占用太高的内存，像象棋和围棋这样的棋类游戏就属于这一类。第二种是我们不知道解决问题的方法是什么(例如人脸识别)，要解决这类问题，我们需要一些全新的东西(例如机器学习)。几乎所有当代的人工智能研究都关注着这两类问题之一。

现在，我们来看看图1中最困难的问题——通用的、人类水准的智能机器这一宏伟梦想。毫无疑问，这一挑战吸引了无数人的关注。而其中最早的以及最有影响力的思想家之一，就是我们的老朋友，艾伦·图灵。

图灵测试

20世纪40年代末至50年代初，第一台计算机的出现引发了一场公开辩论，辩论主题就是这一现代科学奇迹的潜力如何。这场辩论中最瞩目的贡献当归属于一本名叫《控制论》的书，由麻省理工学院数学教授诺伯特·维纳(Norbert Wiener)撰写。这本书将机器与动物大脑神经系统做了详细对比，并触及了许多有关人工智能的想法。《控制论》引起了公众的极大兴趣，但事实上，除了特别专注于此的科学家以及专业数学家，大多数人无法理解这本书。诸如机器是否能“思考”的问题开始在新闻界和广播节目中引起了有模有样的争论(1951年，图灵本人还参加了相关主题的一个BBC广播节目)。虽然还没有名字，但人工智能的萌芽开始浮现。

在公众辩论的推动下，图灵开始认真思考人工智能的可能性。他对公众辩论中经常提到的“机器做不到×××”的说法非常恼火(例如思考、

推理或者进行类似创造性的工作)。他想让那些认为“机器不能思考”的人彻底闭嘴，于是提出了一个测试，现在，我们称之为图灵测试。自1950年第一次提出以来，图灵测试一直具有巨大的影响力，直至如今，它仍然是一个严肃的研究课题。不过，令人遗憾的是，到目前为止，它仍然没能让怀疑者彻底沉默，我们接下来就来讨论原因。

图灵测试的灵感来源于维多利亚时代一种叫作“模仿游戏”的室内游戏。“模仿游戏”的基本玩法是通过对一个人提问，从回答来判断对方是男是女。图灵建议对人工智能采取类似的测试。测试通常是这样描述的：

人类询问者通过键盘与屏幕彼端的“生物”进行交互聊天，询问者事先并不知道对方是人还是计算机程序。交互纯粹以文本的形式进行：询问者键入一个问题，然后对方给予一个回应。询问者的任务是确定对方是人还是计算机程序。

现在，假设被询问的确实是计算机程序，但是经过一段合理的时间，询问者无法准确判断他是在与人还是计算机交互。图灵认为，你就得承认这样的计算机程序拥有类似人类的智能(或者自主思维以及别的称呼)。

图灵的杰出之处是避开了所有存在争议的问题，直指计算机程序是否“真正”拥有智能(或者意识以及其他说法)、程序是否真正有“思维”(或者意识、自主意识什么的)并不是重点，重点在于它能够做到“乱真”，即让测试者无法分辨出程序和真人。这里的关键词在于“无法分辨”。

图灵测试是科学界中标准技术的一个优秀例子，如果你想确认两种事物是相同的还是不同的，就思考一下如何设计合理的测试来区分它们。若是有一种合理的测试，两种事物其中一种能通过，另一种不能，那么你可以宣称它们是不同的。如果不能通过合理的测试来区分它们，那么就不能声称它们是不同的。图灵测试就是用来区分机器智能与人类智能的方法，测试的方式是人类询问者是否能够分辨出与之交流的是机器还是人。

然而在这个问题上我们得谨慎一些，多年来，许多定义人工智能的方式都遇到过类似的困境，它们总是根据所使用的技术方法来定义人工智能。例如，如果你最喜欢的人工智能技术是“时间递归最优学习”[5](我随机挑了个时下最流行的人工智能相关词汇)，那么你可能更倾向将人工智能的挑战定义为能够使用时间递归最优学习方式通过图灵测试，从而

排除掉其他的技术方法。我们需要的是智能行为的测试，它独立于实现智能行为所使用的技术方法之外。图灵测试通过将询问者与测试对象分开来实现这一点：询问者只能通过提问和对方的回答来进行判断——询问者发送问题，并接收测试对象发来的答案。就图灵测试而言，被测试的对象是一个黑盒，从某种意义上来说，我们不允许检查其内部结构，只能对其进行输入和输出。

图灵于1950年在著名国际期刊《心智》上发表的《计算机器与智能》论文中[9]，描述了他的测试。尽管在此前有许多关于人工智能思想的论文，但图灵是第一个从现代数字计算机的角度探讨这个问题的，因此，他的文章被公认为首篇有关人工智能的论文。

毫无意义的图灵测试

图灵测试简单、优雅、易懂，然而很不幸，若是将它奉为人工智能测试的圭臬，则无法回避一个致命问题——哪怕时至今日，它仍然存在。

这个问题在于，大多数试图通过图灵测试的程序员都倾向于使用一些简单的小技巧来“骗过”询问者，使他们相信自己正在与真人打交道，即用小伎俩去迷惑询问者，而并非真正去探寻与解决智能行为的本质。最著名的例子应该是名为ELIZA的程序，它诞生于20世纪60年代中期，由美籍德国人、麻省理工学院计算机科学家约瑟夫·魏岑鲍姆(Joseph Weizenbaum)创造。魏岑鲍姆本人倒是无意让ELIZA挑战图灵测试——ELIZA原本是一项严肃的、富有影响力的心理治疗科学实验——但遗憾的是，自此以后，ELIZA就成了采用肤浅的小伎俩伪装人工智能的代名词，尤其在图灵测试领域。

ELIZA扮演的是一名精神病医生的角色，与病人交谈，提出开放式问题，让病人试图敞开自我。ELIZA学习了一些关键词，比如“家人”“朋友”“孤独”“悲伤”等，然后使用关联了关键词的封装脚本来提出后续问题。明白原理以后，你不用太费劲就能轻易看穿它使用的封装A及镜像反问伎俩。以下是现代版本的一段对话A隐藏编程对象的属性及实现途径，只公布其外部接口进行访问。(ELIZA的回复使用斜体字)[10]：

我很孤独。

你经常感觉到孤独吗？

是的，几乎每天都觉得。

为什么你说几乎每天呢？

我想是家里人的原因。

告诉我一些你家里人的情况吧。

他们讨厌我。

是什么让你相信他们讨厌你呢？

他们对我非常刻薄。

所以你刚才说你很孤独？

生命到底有什么意思？

也许你在生活中可以找到生命的意义。

存在的目的是什么？

是因为你的计划让你思考存在的目的吗？

开场白看起来很不错，但几句对话以后，这种互动的肤浅性暴露无遗——显然ELIZA根本没有理解到对话的意义(当然，魏岑鲍姆也从来没声称它能做到理解对话，这并不是他的实验目的)。

魏岑鲍姆没有预料到ELIZA会在精神病学术界以外引起极大反响，显然他自己都很惊讶，用户对他的简单实验性程序非常沉迷，ELIZA正式进入了人工智能的传说中。20世纪80年代，我在读博士期间，就听说过有孤独症和社会适应不良的研究生，整晚整晚与ELIZA聊天，以取代他们无法建立的正常人际关系。坊间甚至还流传着一个可怕的故事——希望只是虚构的——据说ELIZA无意中鼓励一个抑郁症的学生自杀。魏岑鲍姆要求自己的秘书测试一下是否会出现这样的结果，令他吃惊的是，女秘书要求在她和ELIZA对话期间，魏岑鲍姆必须离开房间，尽管她很清楚这种互动的本质并非真人交流。

ELIZA的影响力，以每年一度的罗布纳奖人工智能竞赛的形式流传至今。也不知道图灵本人能否想象他的测试会有这么多人来参与，但1990年美国富豪发明家休·罗布纳(Hugh Loebner)却真正地支持了这件事情。每年，罗布纳基金会邀请参赛的计算机程序员参与图灵测试，试图让评委们相信与之对话的计算机程序是真人，挑战成功的参赛者将获得十万美金。

罗布纳奖竞赛很快就引起了各种争议。著名人工智能科学家马文·明斯基(Marvin Minsky)认为这是一个“毫无意义的宣传噱头”，这项竞赛仿佛每年都会产生至少一个无可救药的误导性标题。问题就在于，罗布纳竞赛的参赛者很大程度上都是ELIZA的变奏曲，他们并没有试图让询问

者开展一场有意义的对话，充分证实人工智能拥有类人的领悟力、理解力和常识，而是使用一些误导技巧来赢得比赛。下面是一段2009年的获奖作品对话片段，来自名为“进展无限”的人工智能，它由著名人工智能评论员、计算机象棋专家大卫·利维(David Levy)编写(“进展无限”的回答使用斜体字)。

嗨，过得怎么样？

你好，你过得怎么样？

挺好的，你是谁？

这个，我在学习工程学，这应该会让你猜想到我是谁。

不好意思，不太清楚。

我迫切想知道有关更多实验的知识。

你来自哪里？

我不知道，不过别忘了我不可能同时出现在三个地方。

你住在哪个城市？

哪个城市？这个问题挺棘手，问题在于，城市看起来都差不多。

并不是这样啊，你住在哪里？

我想是河流边上的老牛奶厂吧。

真有意思，这么说你是个农民吗？

好吧，如果你询问一个生理学家，他不会认为我是个农民。他会说我的主要成分是水——足够装满一个大啤酒桶——再加上一些铁、磷和许多平淡无奇的成分。

对这么一个简单的问题而言，这真是个奇怪的回答，你觉得呢？

我当然会的，看我的!

看起来比ELIZA复杂得多，但显然离人类之间进行的对话还有非常漫长的道路要走。

ELIZA算是某种互联网产物的始作俑者，一提到这个产物，研究人工智能的程序员们都得抓狂，那就是互联网聊天机器人。它们是基于互联网的程序，试图让用户参与对话，通常活跃于推特等社交媒体平台。开发能够进行有意义对话的程序是一门严肃的研究课题，但绝大多数所谓的互联网聊天机器人都只是使用了关键字联想的脚本而已，再辅以一些类似反问、提出通用问题之类的小伎俩，就如ELIZA一样。因此，它们的聊天内容都是肤浅并且乏味的，这种聊天机器人，并非人工智能。

人工智能的多样性

尽管图灵测试在今天已经没什么用了，但它仍然是人工智能重要的

组成部分，因为它第一次给了对这门新兴学科感兴趣的研究人员一个明确的目标。当有人问起你研究的目标是什么，你可以给出直接并且精准的答案：我的目标是创造一台能够真正意义上通过图灵测试的机器。时至今日我不得不说，除了特别严肃认真的人工智能研究人员可能会给出这样的回答，已经鲜有研究员会以此为目标了。但至少它在历史上有着至关重要的作用，我也相信，如今它仍然具备一定的重要性。

图灵测试最吸引人的地方无疑是它非常简单明了，尽管看起来很清晰，它仍然对人工智能提出了许多挑战。

假设你是图灵测试中的询问者，你能够判断跟你交流的对方是真人，其中很重要的依据就是：另一端的测试对象展示了对你提出问题的理解，并且能够给出人类应该给出的答案，此时你发现跟你交流的其实是一段人工智能程序，这才算通过。现在，就图灵测试本身而言，问题是能够解决的。程序正在做一些与人类相同的事情：结束争辩。但是，其中存在两种逻辑上截然不同的情况：

1. 程序确确实实理解了与询问者的对话，这种理解与人类的理解大致相同。
2. 程序并没有真正理解与询问者的对话，但可以模拟出理解对话以后的回复。

这两种情况简直是天差地别，宣称能做到情况1——程序真正能理解对话的，明显比只能做到情况2厉害得多。情况2只需要我们创建表现出能理解对话的程序即可。

我想大多数人工智能研究人员——可能还包括大多数本书的读者——都能轻松理解情况2是容易可行的(至少原则上是可行的)。但要他们接受情况1是可行的这一点上，恐怕得需要更多的说服力。事实上，我们不知道如何证明一段程序如情况1所描述那样运行——图灵测试本身也没有这么要求(我怀疑，图灵本人肯定会对这两者的区别表示恼火：他会指出，发明图灵测试的主要原因之一就是为了结束关于这两者区别的争论)。因此，立志构建一个类似情况1的程序，比起情况2的程序而言，更具有雄心壮志，也更具有争议性。

构建出具有人类的理解力(或者说是意识之类的词)的目标程序，被称为强人工智能，而次一级的目标，即构建虽然没有具备人类的理解力，但是可以模拟出特定能力的程序，被称为弱人工智能。

超越图灵测试

以图灵的“无法分辨”为原则，图灵测试有许多变种。例如，在一个更强大的测试版本中，我们可以想象机器人试图把自己伪装成人类进行日常生活。在这里，“无法分辨”是以一种非常苛刻的方式来诠释的——这意味着机器人与人类完全无法区别(当然，你不能去解剖它们，所以本质上还是黑盒测试)。在可预见的未来，这类情况只能时常出现在小说里面。事实上，一个无法区分机器人和人类的世界，就是绝佳的电影及小说创作素材。以此为基础的电影可不少，1982年，雷德利·斯科特(Ridley Scott)的经典作品《银翼杀手》中，年轻的哈里森·福特(Harrison Ford)饰演的主角每天都在进行神秘的测试，目的是确认漂亮的年轻女性实际上是机器人。类似的主题还有许多，比如2014年的《机械姬》。

虽然《银翼杀手》中的场景还不曾出现，但研究人员已经在探究图灵测试是否存在一些变式，可以有意义地测试真正的智能，并且可以破解类似网络聊天机器人那种骗过询问者的诡计。一种非常简单的想法就是测试理解力，就如著名的威诺格拉德模式都是一些需要理解力的简短句子，下面举例说明[11]：

> 句子1a：市议员拒绝给示威者颁发许可，因为他们担心暴力。
>
> 句子1b：市议员拒绝给示威者颁发许可，因为他们鼓吹暴力。
>
> 问题：谁担心/鼓吹暴力？

注意，这两个句子中只有一个词语不同(加下画线词语)，但这个小小的不同会影响整个句子的含义。测试的重点在于确定在每种情况下，被测试对象都能准确地辨认出“他们”指代的是谁。在句子1a中，“他们”显然是指市议员(市议员担心示威者出现暴力行为)，而在句子1b中，“他们”指的是示威者(市议员担心示威者会鼓吹暴力行为)。

这里我们再举一个例子：

> 句子2a：奖杯不能放进棕色手提箱里，因为它太大了。
>
> 句子2b：奖杯不能放进棕色手提箱里，因为它太小了。
>
> 问题：什么东西太大/太小？

显然，在句子2a中，是奖杯太大，句子2b中，则是手提箱太小。

大多数稍有文化和常识的成年人都能轻易辨认出两个例子中的区别，类似这样的问题完全难不倒真人，但类似聊天机器人使用的那些小

伎俩就会被这种简单的句式辨别出来。为了给出正确答案，被测试者需要真正理解这些句子，并且要对讨论的场景有一定了解。比如，要辨别句子1a和1b的差异，你需要了解什么叫示威者(示威往往有可能导致暴力行为)，什么叫市议员(他们有权批准或拒绝示威许可，并且他们会努力避免暴力行为出现)。

人工智能所面临的另一个类人挑战是理解人类世界，以及支配其中的许多约定俗成的规则。下面是心理学家兼语言学家斯蒂芬·平克(Steven Pinker)设计的一段简短对话：

> 鲍勃：“我要离开你。”
>
> 爱丽丝：“她是谁？”

你能解释这段对话的意思吗？当然，对你来说这太容易了，这就是肥皂剧中经常会出现的场景：鲍勃和爱丽丝在谈恋爱，现在鲍勃的话让爱丽丝认为他移情别恋了，她想知道那个插足他们感情的女人是谁。并且，我们可以推测出，爱丽丝一定很生气。

但是你怎么能让计算机编程来理解这样的对话呢？这样的理解力对于听故事、写作来说，都是必不可少的。如果能理解这种贯穿于生活中的常识性的东西，以及人类的信仰、欲望和关系，计算机程序就可以欣赏类似《东区人》的肥皂剧了。我们都拥有这样的能力，这似乎也是分辨强人工智能和弱人工智能的关键点。对于如何让机器拥有这样的理解力，我们还只有模糊的想法，到目前为止，没有一个成功的案例。机器要想理解这样的场景，回答关于这些场景的问题，还有漫长的道路要走。

通用人工智能

虽然如我们所证明的那样，人工智能的宏伟梦想听起来简单明了，但要想实现它，目前研究人员简直是毫无头绪，连给它下一个精准的定义都难上加难，更不知道什么时候才能做到。因此，尽管强人工智能是人工智能故事中一个重要而迷人的部分，但很大程度上，它与当代的人工智能研究无关。去参加一个尖端人工智能研讨会，你几乎听不到有关内容的讨论，倒是在深夜的酒吧里面能听到不少。

另一个退而求其次的目标就是制造具有普遍人类智能水准的机器，现在，通常称其为通用人工智能(Artificial General Intelligence，

AGI)， AGI大致等同于一台拥有一个普通人所拥有的全部智慧能力的计算机，包括使用自然语言交流(参见图灵测试)、解决问题、推理、感知环境等能力，与一个普通人处于同等或者更高等级的智能水准。一个能够实现的AGI系统将能完成图1所示的所有任务，甚至更多。关于AGI的文献通常不涉及自我意识或者自主意识之类，因此AGI被认为是弱人工智能的弱版本[12]。

然而，即使这个次一级的目标，也是当代人工智能研究的边缘。与之相反，人工智能领域的研究人员更关注的是构建能够执行当前必须依靠人脑来执行的任务的计算机程序——逐步遍历图1所示的任务集。这种人工智能的实现方法——让计算机完成某种特殊任务，有时候被称为狭义人工智能。但这个说法并非人工智能科研领域的术语。事实上，如果你在某个人工智能大会上使用这种方式表达，别人会认为你是个外行。我们不会把目前所研究的人工智能加个“狭义”的前缀，因为所谓的狭义人工智能，就是人工智能。对于那些渴望机器人全能管家的人而言，这不是个好消息。但对那些担心机器人叛乱统治人类的人而言，这可能还真是个好消息。

所以，现在你应该明白，人工智能究竟是什么东西了，并且知道它为何如此难以实现。但是，人工智能的研究人员究竟是怎么做到解决各种问题的呢？

心智或者大脑？

我们怎样才能够让计算机产生人类水准的智能行为呢？历史上，人工智能研究人员对这个问题有两种实现方式。粗略地说，第一种是试图建立思维模型：有意识的推理、认知、解决问题的过程，我们在生活中都会用到的过程。这种方法被称为符号人工智能，因为它使用各种符号来代表各种事物，并对此进行推理、认知等行为。例如，机器控制系统中的符号“room 451”可能是机器人用来指代你卧室的符号，而“cleanroom”则是用来定义房间清洁这一活动的符号。当机器人确定要做什么的时候，它会明确使用这些符号。例如，当机器人决定执行“cleanroom(room 451)”操作，这就意味着机器人要对你的卧室进行清洁。符号就意味着机器人语言中的事物及行为。

从20世纪50年代中期到80年代末，30多年的时间里，符号人工智能一直是构建人工智能体系最流行的方式。它有许多优势，或许最重要

的一点是，它的过程是透明的：当机器人认为它应该执行“cleanroom(room 451)”操作时，我们可以理解为它知道即将做什么。不过，我认为符号人工智能之所以这么流行，是因为它反映了我们有意识的思维过程。我们本身“思考”的方式就是用符号或者文字流程化了，在决定做什么之前，我们可能会跟自己来一场心灵对话，讨论各种方案的利弊，最终决定执行某一种。符号人工智能渴望捕捉这一切，我们将在第二章讨论相关问题。在20世纪80年代初期，符号人工智能的发展达到巅峰。

另一种模拟智慧的方式是模拟大脑，这是一种极端的可能性，即试图在计算机中模拟一个完整的人类大脑(也许还得包含完整的神经系统)。毕竟，人类的大脑是我们唯一确定能产生智慧的事物。这种方法的问题在于，大脑是一个复杂得难以想象的器官：人脑包括大概1000亿个相连的神经元，我们对它们的成分、结构和如何运作的了解还达不到复制大脑结构的程度。这不是一个很快能够实现的目标，甚至我怀疑它根本没有可实现性(我不得不遗憾地说，这也无法浇灭某些人尝试进军该领域的热情)[13]。

不过我们可以从大脑结构中获取一些灵感，并以此结构为基础构造智能系统中的组件。这个研究领域被称为神经网络——这个名字来自我们大脑微观结构中细胞信息处理单元神经元，它们是呈网状连接的。神经网络的研究可以追溯到人工智能出现之前，并沿着人工智能的主流研究发展。正是神经网络研究在21世纪取得的突破性进展，才带来了目前人工智能研究领域的繁荣。

符号人工智能和神经网络人工智能是两条截然不同的道路，所使用的方法也完全不同。在过去的60年里，它们都有着各自的辉煌和没落，我们也将在后续章节讨论到，两种流派彼此之间也爆发过激烈的争论。然而，在20世纪50年代，人工智能作为一门新兴的学科诞生，符号人工智能在很大程度上占据了主流地位。

[1] 《弗兰肯斯坦》被认为是世界上第一部真正意义上的科幻小说。

[2] 查尔斯·巴贝奇发明的分析机被认为是最早期的计算机雏形，而阿达·洛芙莱斯的算法则被认为是最早的计算机程序和软件。

[3] 图灵指出的悖论即著名的“停机问题”，如果存在图灵机A，对它输入任意图灵机，都能够判定其运行结果是“停机”或者“不停机”，那么可以构造图灵机B，调用图

灵机A但输出永远与之相反，即B的输入经A判定“停机”，则B“不停机”，如果B的输入经A判定“不停机”，则B“停机”。那么，图灵机B的输入为图灵机B本身时，则出现矛盾。B只有在“停机”的时候才能“不停机”，反之亦然。此悖论与理发师悖论类似，理发师声称“给而且只给那些不给自己理发的人理发”。如果理发师不给自己理发，那么根据定义，他要给自己理发；如果理发师给自己理发，那么根据定义，他不能给自己理发。

[4] 1英尺≈0.30米。

[5] 2016年NIPS(神经信息处理系统大会，是一个关于机器学习和计算神经科学的国际会议)上，出现一家神秘的创业公司：火箭人工智能，宣称要基于一项革命性的“时间递归最优学习”技术来进行人工智能项目开发。短短几小时，吸引了五家风投公司对此项目进行评估，估值高达上千万美元。事实上，这家公司是参会的专家们注册来恶搞的，该技术纯属子虚乌有，专家们旨在以此来讽刺人工智能领域的投资泡沫太过严重。

第二部分 我们是怎么走到这一步的

第二章 黄金年代

图灵在他发表的论文《计算机器与智能》中介绍了图灵测试，为人工智能学科迈出第一步做出了重大贡献。只是在当年，这个划时代的贡献显得颇为孤独。那时候，人工智能还没有命名，没有形成一门学科，也没有研究团队。彼时图灵的贡献都只是概念性的，比如图灵测试——因为人工智能系统还未出现。仅仅过了10年，到了20世纪50年代末，一切都发生了天翻地覆的改变：新的学科建立，并有了专属的名称，研究人员可以自豪地展示用以揭示智能行为基础组成部分的第一个试验性系统。

接下来的20年是人工智能的第一个繁荣期。1956年到1974年间，人工智能研究领域掀起了乐观积极、迅速发展的浪潮，这一时期被称为人工智能的黄金年代。在这个年代里，研究人员不知失望为何物，一切皆有可能。这一时期构造的人工智能系统，堪称人工智能史上传奇的经典标准。它们拥有奇奇怪怪的名称，比如**SHRDLU**、**STRIPS**或者**SHAKEY**——都是简短的名字，很多来源于首字母缩写，全用大写字母表示。这大概是源于当时计算机文件名的限制(以这种方式命名人工智能的习惯一直沿用至今，尽管已经毫无必要)。按照现代标准，用来建立这些人工智能的计算机运行速度慢得难以想象，内存也小得难以置信，使用起来也麻烦得难以忍受。今天我们开发软件理所当然会使用到的那些工具，当时并不存在，事实上，即使存在，也没法在当时的计算机上运行。那时候“黑客”文化已经悄悄抬头，人工智能研究人员大多在晚上工作，因为他们需要的计算机资源在白天要用在更重要的工作上。他们还必须发明各种巧妙的编程技巧[14]，让复杂的程序能在当时简陋的计算机上运行——许多技巧后来成为各种技术标准，它们起源于20世纪60到70年代的人工智能实验室，现在的人们只能模糊地回忆起来——如果还有人记得的话。

但到了20世纪70年代中期，人工智能的研究开始停滞不前，除了最初的简单试验，并没有多少有意义的进展。这门初生的学科几乎被研究资助者扼杀于摇篮中，科学界也逐渐开始相信，人工智能这种被寄予了

无限希望的新生儿，实则前途渺茫。

本章将回顾人工智能发展的前20年，我们将研究在此期间构建的一些关键系统，并讨论当时在人工智能领域最重要的技术之一：被称为“搜索”的技术。即使时至今日，它仍然是诸多人工智能系统的核心组件。我们还会接触到一个抽象的数学理论——计算复杂性，这是在20世纪60年代末到70年代初建立并发展起来的理论，用以解释为什么人工智能所面临的诸多问题从本质上来说难以解决。计算复杂性给人工智能的发展蒙上了一层阴影。

我们将从公认的黄金年代起点开始：1956年的夏天，一位名叫约翰·麦卡锡(John McCarthy)的年轻美国学者，为这一领域命了名。

人工智能元年的盛夏

麦卡锡生活在美国现代科技创建的年代，在20世纪50到60年代，他以近乎不可能的才华，为计算机领域贡献了一系列新的理念。这些理念在现代被认为理所当然，以至于我们很难想象当年如果不是因为这些理念，计算机科学就不可能有进一步发展。他最著名的成就应该是被称为LISP的编程语言，几十年来，LISP一直是人工智能研究人员首选的编程语言。虽然编程语言一向是难以阅读的，即使以我这种专业程度为标准来看，LISP语言也是个怪胎。因为在LISP语言中，(所有(程序(看(起来(都(是(这样的)))))))。一代又一代的程序员总是开玩笑说LISP就是“一堆毫无关联的傻括号”[15]。

麦卡锡在20世纪50年代中期发明了LISP语言，令人惊讶的是，近70年后，全世界仍然在教授和使用它(反正我每天都要用)。想一想吧，麦卡锡发明LISP语言时，美国总统还是德怀特·戴维·艾森豪威尔，尼基塔·谢尔盖耶维奇·赫鲁晓夫还是苏联共产党总书记；在中国，毛泽东主席正在如火如荼地领导第一个五年计划，全世界电脑总数才几台。而时至今日，LISP语言竟然还在广泛使用中!

麦卡锡出生在波士顿的一个移民家庭，在很小的时候就展露出非凡的数学天赋。从加州理工学院数学专业毕业以后，这个20岁出头的小伙子被任命为新罕布什尔州达特茅斯学院的副教授。麦卡锡在加入达特茅斯学院之前就对计算机产生了浓厚兴趣，1955年，他向洛克菲勒研究所提交了一份提案，希望获得资金来在达特茅斯学院开办一所暑期学校。

如果你不是一名学者，成年人的暑期学校听上去似乎还挺奇怪的。不过即使在今天，这也是一个久负盛名且受人喜爱的学术传统，主要目的是召集世界各地有共同兴趣的研究人员，让他们有机会在一段时间里见面交流及共事。之所以选择暑期，显然是因为一年的教学任务已经结束，学者们大多可以卸下沉重的教学压力，轻松一下。当然啦，既然选择在如此迷人的地方组织暑期学校，必要的社交活动也是不可缺少的。

对一所令人难忘的暑期学校来说，另一个不可或缺的要素就是一张星光四射的学员名单。当时的人们可能没预见到，达特茅斯暑期学校会集了未来几十年内在人工智能领域起决定性作用的大部分关键人物。不过邀请名单上还有一个特别令人沉痛的名字：普林斯顿大学数学家小约翰·福布斯·纳什(John Forbes Nash， Jr)。他于6年前刚刚取得博士学位，在一篇论文(只有短短28页)中引入了一个叫作“非合作博弈”的概念。纳什提出的概念在随后的几十年里成为经济理论的基石，并最终在1994年荣获诺贝尔经济学奖。但纳什本人并没有享受到他这项工作的乐趣：就在获得博士学位的几年后，他就被妄想症和偏执狂等精神疾病所吞噬，这使得他在之后几十年都远离了学术生活。不过令人高兴的是，他凭着惊人的毅力恢复良好，最终在1994年获奖。他的故事被写成了书籍《美丽心灵》，同名电影于2002年荣获奥斯卡金像奖最佳影片奖等多项殊荣[16]。

达特茅斯暑期学校名单令人好奇的原因还有一个。除了学术界的大佬们，暑期学校还接待了来自工业界、政府及军方的代表(甚至还有兰德公司——这家以加州智囊团为班底的智库公司，在20世纪60年代因为毫无感情地争辩如何“赢得”核战争而臭名昭著。而仅仅在十来年前，学术界、工业界、政府和军方联合起来执行曼哈顿计划，制造出了全世界第一颗原子弹——这也是美国科技实力的明确体现)。这种联合体——学术界、工业界、政府和军方联合，是美国在第二次世界大战后几十年里计算机技术发展的特色，也是美国在未来60年内确立人工智能领域国际领先地位的核心。

1955年，麦卡锡向洛克菲勒研究所撰写计划书申请经费的时候，他不得不给这项活动起个名字，他选择了“人工智能”这一名称。麦卡锡对他的活动抱有不切实际的高度期待——这也成了人工智能领域令人生厌的传统——他高呼：“我认为我们可以取得重大进展……如果这些经过精挑细选的科学家能在一起共事一个夏天的话[17]。”

到暑期学校结束之时，共事的科学家们也没取得什么重大进展，但麦卡锡选择的名字却流传下来，逐渐形成了一门新生的学科。

不幸的是，麦卡锡选择的这个名字在很长的时间里受到了无数诟病。首先，人工智能中的“人工”一词，英文为“artificial”，它还有一层意思是“虚假”“赝品”——谁会乐意研究虚假的智能呢？另外，“智能”(intelligence)这个词汇来源于“智力”(intellect)，而事实上，自1956年以来，人工智能界一直在努力完成的诸多挑战，人类去做的时候似乎并不需要太高的智能。相反，正如我们在上一章所讨论的，过去60多年中，人工智能一直致力于解决的许多最重要也是最困难的问题，似乎一点都体现不出智能——这样的事实一再让那些新入该领域的人感到惊愕和困惑。

不过，“人工智能”这个由麦卡锡选择的名字，至今仍然沿用。那些参与暑期学校的学者，以及他们的学术继承人，多年来不断地致力于人工智能领域的研究，直至今日。那年夏天，是公认的人工智能学科诞生元年，当时，它听上去前途无量。

达特茅斯暑期学校之后那些年是人工智能令人兴奋的成长期，并且，在某段时间里似乎处于飞速发展状态。几十年里，四名暑期学校的成员主导着人工智能的发展。麦卡锡本人在斯坦福大学创立了人工智能实验室，位于现在的硅谷中心地带，马文·明斯基在马萨诸塞州剑桥市的麻省理工学院也创立了一个实验室，艾伦·纽维尔(Alan Newell)和他的博士生导师赫伯特·西蒙(Herbert Simon)则去了卡内基-梅隆大学。这四位天才带领着他们的学生建立的人工智能系统，是我们这一代人工智能研究工作者的图腾。

然而，黄金年代也有着幼稚、盲目乐观的特点，研究人员对这个领域可能的发展速度做出了鲁莽的预测，令人工智能的实际发展受到阻碍。到了20世纪70年代，美丽的梦幻结束了，一场恶性循环开始——人工智能的繁荣和萧条不断上演，在未来几十年内反复交替。但是，不管历史如何去批判这一时期，我都满怀喜爱之情去思考这一时期涌现出的人物，以及他们所做出的伟大成就。

分治策略

正如我们所知，通用人工智能是一个庞大而模糊的目标，很难直接

达成。因而，黄金年代采用的策略是分而治之。也就是说，与其一口气尝试构建完整的通用智能系统，不如识别出通用人工智能系统所需的各项不同能力，分别构建具备这些能力的体系。这一概念隐含的假设是，如果我们能够成功构建通用人工智能所必需的每一种智能系统，那么，以后将它们组合成一个整体，比起直接构建完整的通用智能系统简单得多。这一基本假设——通用人工智能的研究方向是将精力集中在解决智能行为的各种组件上——成为人工智能研究的标准方案。人们争先恐后地建造各种能够表现智能行为能力的组件。

那么，研究人员关注的主要问题有哪些？首要的也是最难解决的问题，又偏偏是我们认为理所当然应该具备的：感知。一台机器要在特定的环境中智能地工作，就必须能够获取周围环境的信息。我们人类通过各种机制感知世界，包括五种感觉：视觉、听觉、嗅觉、触觉和味觉。因此，就得研究制作能够感知外界的传感器。如今的机器人使用各式各样的人工传感器来接收有关环境的信息——雷达、激光雷达、红外线测距仪、超声波测距仪等。但是制造这些传感器本身就是一项复杂的工程，但这还只是问题的一部分。不管数码相机的光学系统有多好，不管相机的图像传感器有几百万还是几千万像素，最终相机所做的事情只是把看到的图像解析成一个个网格，然后为网格中的每一个单元格分配数字，表示它的颜色和亮度。所以，哪怕配备了质量最好的数码相机的机器人，最终它收到的信息也只是一连串的数字。因此，感知领域的第二个挑战是如何解析这些原始的数字，理解它所“看到”的图像。这项挑战远比制作传感器难得多。

通用智能系统的另一个关键能力是通过学习积累经验，这引出了一个名叫机器学习的人工智能研究分支。就像“人工智能”这个名字本身一样，机器学习也是一个极其不恰当的术语，听起来像是一台机器自己在引导自己学习，从无知开始，变得越来越聪明。事实上，机器学习跟人类学习是完全不同的，它的学习指的是解析和预测数据。例如，在过去数十年里，机器学习的一个巨大成就是出现了可应用的人脸图像识别技术，通常的实现方法是给予程序诸多希望它学习的例子。因此，通过许多人脸图片以及对应人名来训练人脸图像识别程序，最终达到的效果是给它呈现一张图片时，程序就能正确地给出所示图片对应的人名。

问题解决和制订计划是两个相关联的能力，它们也与智能行为有关。它们都要求用给定的动作组合来达成目标，而实现的关键在于找到

正确的动作序列。例如，象棋或者围棋之类的棋盘游戏，目标是赢得比赛，行动是在棋盘上移动(或放置)棋子，挑战是要找出正确的下棋步数以达到赢得比赛的目的。而如我们所了解的，在解决问题和制订计划中最根本的难点在于，只需要考虑移动棋子的所有可能结果，虽然从原理上看起来很容易做到，但是实际上执行起来却非常困难，因为棋子可能移动的步骤实在是太多了。

推理也许是诸多与智能相关的行为中最令人崇拜的一种：它基于现有的事实，用强大的逻辑方式获取新的知识。举一个著名的例子，若你知道“所有的人都是凡人”，你也知道“迈克尔是一个人”，那么就可以推理得出“迈克尔是凡人”的结论。一个真正的智能系统很容易做到这一步，然后，利用新的知识做出进一步推理。例如，知道“迈克尔是凡人”之后，就能推论出“迈克尔有朝一日会死去”“迈克尔死去以后，将不会复活”等等。自动推理就是赋予计算机这样的逻辑推理能力。在第三章中，我们将看到，长期以来，人们都认为这种推理能力应该是人工智能的主攻方向。虽然现在已经不是主流了，但自动推理仍然是人工智能领域中一个重要的分支。

最后是自然语言理解，这涉及让计算机理解人类的语言，比如中文或者英文。目前，计算机程序员为计算机编程时，必须使用人造的机器语言：某种有着精确定义、明确规则、专门为计算机构造的语言。这些语言(目前最著名的应该是Python、Java和C语言)比起自然语言(如中文或英文)简单得多。很长一段时间以来，让计算机理解自然语言的主要方法是试图为自然语言加以精准的定义和明确的规则，就像我们定义计算机语言一样。而事实证明，这是不可能的。自然语言太过灵活、模糊和易变，无法用这种方式严格定义，而语言在日常生活中的使用方式更是阻碍了对其进行精确定义的尝试。

SHRDLU和积木世界

SHRDLU系统是黄金年代最受欢迎的成就之一(这个奇怪的名字来源于当时印刷机上的字母排列——程序员就喜欢弄点晦涩的笑话)。1971年[18]，斯坦福大学博士生特里·威诺格拉德(Terry Winograd)开发了SHRDLU系统，旨在展示人工智能的两个关键能力：解决问题和自然语言理解。

SHRDLU的问题解决组件是基于人工智能界最著名的实验场景之

一：积木世界。积木世界是一个包含了许多彩色物体(方块、盒子和锥体)的模拟环境。使用模拟环境而不是真实环境去构建机器人，初衷在于将问题的复杂性降低到可管理的程度。在积木世界中，SHRDLU的问题解决模块可以根据用户的指令来排列对象，也可以使用模拟机械手臂来操作对象。我们可以在图2中看到积木世界的初始状态和目标状态。人工智能要完成的挑战是如何将初始状态转换为目标状态。并且，你只能使用固定的操作集合来完成状态转换：

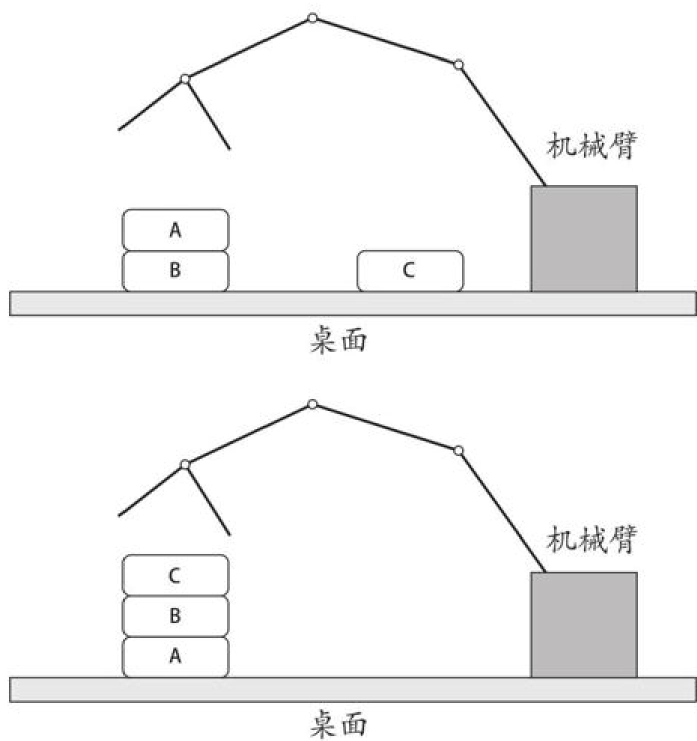


图2　积木世界上图是初始状态，下图是目标状态，你要怎么转换呢？

·从桌面拾取对象x。此处指的是用机械臂将对象x(可能是积木或者锥体)从桌面上拿起来，只有当同时满足对象x在桌面上，以及机械臂当前为空的时候，才能执行此操作。

·将对象x放在桌面上。只有当机械臂携带对象x的时候，才能执行此操作。

·从对象y上拾取对象x。只有满足机械臂当前为空，对象x必须在对象y的上方，并且对象x上方没有其他对象的时候，才能执行此操作。

·将对象x放在对象y的上面。只有当机械臂携带对象x，以及对象y的顶部没有其他

物品的时候，才能执行此操作。

积木世界里的一切行为，最终都会简化为这四个操作步骤，且只能执行这四种操作。所以，要实现图2中的状态转化，可以先按照如下步骤执行：

·从对象B上拾取对象A

·将对象A放在桌面上

·从桌面上拾取对象B

上述行动完成以后，积木世界会是什么状态？你能补写出剩下的操作步骤吗？(当然，这不难，但是挺无趣。)

积木世界可能是整个人工智能领域中研究得最多的场景，因为在积木世界中，机械臂完成拾取物品并移动它的操作，听上去与现实世界中我们为机器人设想的任务类似。但是，在SHRDLU(以及诸多后续研究)的积木世界，如果要作为开发使用人工智能技术的场景，还有着严重的局限性。

首先，积木世界是一个封闭的世界，这意味着引起世界变化的唯一因素就是SHRDLU。这就像你一个人住的时候可以笃定地相信，你睡觉前把房门钥匙放在哪里，醒来以后它仍然在原处。如果你家里还有其他人，就会存在别人拿走你钥匙的可能。因此，当SHRDLU把对象 x放在对象 y上面之后，就可以笃定地认为对象 x一定位于对象 y的上方，除非它对相关对象进行过操作。而现实世界并非如此，人工智能系统不能假设自己是世界上唯一的行动者，这样的假设只会让运行结果错误百出。

其次，也是更重要的一点，积木世界是模拟世界，SHRDLU并没有真正操作一个机械臂来拾取对象并移动它们——它只是作为程序假设这么做而已。它模拟了一个世界，并模拟出自己的行为对这个世界产生的影响。SHRDLU从来没有构建一个基于真实世界的模型，也没有检查模拟世界的模型是否和现实世界相匹配，这就是一个极简的假设而已。积木世界是否忽略了机器人在现实世界中运行所面临的绝大多数困难，这也是一个后来研究人员争论颇多的话题。

为了更好地理解这一点，我们以“从对象 y上拾取对象 x”为例，考虑它的操作。站在SHRDLU的角度，这就是一个简单、明了的动作：机器人只需要径直从指定的目的地拿起指定的对象就算执行完成，无须考

虑这个动作涉及的其他东西。因此，程序只需要找到执行任务所需的正确方法和步骤顺序，就能“控制”整个过程，它不会考虑在执行任务的过程中所遇见的各种麻烦事。但想想现实世界，假设机器人在仓库的环境里，那么它就必须识别出哪个是对象 x，哪个是对象 y，这才能成功执行这条命令。好吧，即使解决了对象识别的问题，单单抓取这最后一步，也不是件简单的事情——让机器人在现实世界中操作哪怕是最简单的物件，困难都远比想象的大，即使在今天都是一个富有挑战性的难题。1994年，我以一个年轻学者的身份前往西雅图参加美国人工智能协会会议时，就对这个问题有了一定的了解。我仍然清晰地记得那些参加“清理办公室”竞赛的机器人，其不称职的程度简直令人惊讶。顾名思义，竞赛的目的是制造一个机器人，它可以在类似办公室的环境中安全地移动，识别和收集垃圾——这个环境就类似于现实版的积木世界。此时已经是SHRDLU问世20年后了——结果，哪怕是在比赛中表现最好的机器人，它的执行速度也慢得令人发指，让人差点以为无法完成任务。当然，问题不在于机器人，而在于人们天真地低估了这些问题的难度。事实上，我们错误地以为，某些人类执行起来轻而易举的任务，机器人执行起来也应该同样轻松。

因此，积木世界之所以引起人们关注，是因为它直观且易于理解。看上去它构造出了类似现实世界的机器人工作环境，其实都是我们自己臆想出来的。深入研究一下就能发现，积木世界所研究和解决的，只是用一系列可执行的动作来解决问题的方法。虽然这确实是一个值得关注的领域，但这只解决了机器人在现实环境所面临的问题中极小的一部分，并且绝不是最困难的那部分。

SHRDLU最为人称赞的特性之一，是用户可以向系统发送指令，并且能从系统中接收到类似自然语言的反馈。下面是一段人类与SHRDLU的对话(SHRDLU的话用斜体字表示)：

拾取一个大的红色方块。

好的。(执行命令)

抓取一个锥体。

我不明白你指的是哪个锥体？

寻找一个比你拾取的方块更高的物块，并把它放进盒子里。

我想你所说的“它”指的是比我拾取的方块更高的那个。

好的。(执行命令)

盒子里面有什么东西？

蓝色的锥体和蓝色的方块。

蓝色的锥体由什么支撑？

由盒子支撑。

有几个积木不在盒子里？

四个。

至少有一个比我叫你拾取的那个更窄吗？

是的，红色立方体。

它有支撑吗？

有，桌子。

桌子能拾取方块吗？

不能。

一个锥体能被一个方块支撑吗？

能。

这段对话远比ELIZA有意义得多——这个系统更加自然、丰富，能够处理复杂结构的句子(比如“至少有一个比我叫你拾取的那个更窄吗？”)以及明白代词“它”在对话中指代的具体对象[19]。正是SHRDLU系统的类似自然语言的交互方式，让它在20世纪70年代初受到了热烈的追捧。然而，后来人们发现，SHRDLU之所以能够产生如此丰富的对话，仅仅是因为它处于一个非常受限的环境——积木世界。虽然不像ELIZA那样事先预设了各种处理对话的模式，但它的对话发生场景是受限的，所以对话内容的局限性非常明显。这个系统首次出现时，人们希望它可以提供一条通向处理自然语言的道路，然而，这个希望落空了。

我们这些身在50年后的人很容易发现SHRDLU的局限性，但无论如何，它仍然是一个影响巨大的里程碑式的人工智能系统。

机器人SHAKEY

机器人与人工智能一直密切相关——尤其对于媒体而言。1927年，弗里兹·朗(Fritz Lang)在其导演的经典电影《大都会》中塑造的机器人形象，成为后世无数机器人荧幕形象的模板：长着一个脑袋、两只胳膊、两条腿的类人的机器，再加点儿冷酷凶残的性情。即使在如今，似乎大众媒体上每一篇有关人工智能的文章都得配上机器人的图，长得就像《大都会》里机器人的直系后代。机器人，尤其是类人的机器人，成为人工智能的标志，这并不让人感到惊奇。毕竟，最能体现人工智能梦想的，就是有长得跟我们差不多、拥有类似我们智力的机器人，与我们同

吃同住同劳动。再说了，我想大多数人都会乐意拥有一个机器人管家来帮忙处理生活中的各种问题。

不过在黄金年代，机器人只是人工智能故事中占比相对较小的一部分。原因相当简单：制造机器人又昂贵又费时，坦率地说，还很困难。一个在20世纪60或70年代独立工作的博士生，绝对承担不起建造一个研究级的人工智能机器人的费用。它需要一个完整的研究实验室、专业的工程师、一整套生产车间流水线、专业的生产设备等。另外，能够驱动人工智能系统的计算机太过庞大和沉重，机器人无法携带。所以，对研究人员而言，构建一个类似SHRDLU的程序要比构建一个现实的、可操作的机器人简单得多，也经济得多——反正系统的复杂性和混乱性是能满足现实需求的。

不过，尽管实体机器人的研究很少，但在黄金年代，还是有一个光辉灿烂的人工智能机器人实验成果：在1966年至1972年间斯坦福研究所开展的SHAKEY项目。

SHAKEY是人类第一次认真尝试构建可移动的、实体的机器人，它可以在现实世界完成各项任务，并且自己想出完成这些任务的方法。要做到这些，SHAKEY需要感知所处环境，了解自己身处的位置和周围的状况；还要能接收任务，并自己制订完成任务所需要的步骤；然后按步骤执行任务，同时确保在执行过程中一切顺利，达到预期效果。它所需要完成的任务主要是在类似办公室这样的环境里移动各种盒子。听上去跟SHRDLU很像，不过SHAKEY可不是SHRDLU那样虚拟的系统，它是一个真实的机器人，能够真真正正地操作物体，这可是一项伟大的挑战。

要想达成目标，SHAKEY要集成好多令人望而生畏的智能系统。首先，建造它的工程师得解决一个大难题：开发人员要自行建造机器人，它必须足够小，足够灵活，才能在类似办公室的环境里移动。还得拥有足够强大和精准的传感器，使机器人能够了解周围的情况。为此，SHAKEY配备了一个电视摄像机和激光测距仪，用来确定它和各物体之间的距离。为了探知障碍物，它还配备了一个名叫“猫须”的碰撞探测器。然后，SHAKEY必须拥有在环境中导航的能力，它还能够制订执行任务所需要的步骤。为此，开发人员设计了一个名为STRIPS(斯坦福研究所问题解决系统Stanford Research Institute Problem Solver的缩写)[20]的系统，现在，人们公认STRIPS系统是人工智能规划技术的鼻祖。

最后，所有的智能系统必须流畅、完美地彼此配合，协同工作。任何人工智能研究人员都会告诉你，上述的所有系统，能成功实现其中一个，都是攻克了大难关；若是能让它们作为一个整体工作，其难度会直线跃升好几个数量级。

当然，听上去SHAKEY很神奇，不过它也充分暴露了当时人工智能的局限性。为了让SHAKEY正常工作，设计者不得不大大简化机器人所处的环境，还要降低任务难度。比如，SHAKEY解析它自带的电视摄像机的数据的能力非常有限，几乎只能用来探测障碍物。即便如此，它所处的环境都必须经过特别的粉刷，还需要精心的照明。因为电视摄像机功率太大，所以只在有需要的时候才能打开，打开电源后10秒左右才能产生可用的图像。当时的开发人员一直在与计算机的局限做斗争：比如SHAKEY需要花费15分钟的时间才能设计好怎么完成一项任务，在此期间，它像个傻瓜一样站在那里，一动不动，与周围环境完全隔绝。由于能够完成SHAKEY相关软件运算的计算机体积都太大，重量也太重，所以得用无线电把SHAKEY连接到一台操控计算机上[21]。总之，SHAKEY这种机器人没有任何实用意义。

SHAKEY可以说是第一个成为现实的自动移动机器人，它开创了一系列令人惊叹的人工智能新技术，和SHRDLU一样，由于这些成就，它理应在人工智能历史中获得殊荣。但是SHAKEY的局限性，证明了人工智能离实用的、拥有自主能力的机器人梦想有多遥远，完成这一挑战有多艰巨。

问题解决与搜索

解决问题的能力无疑是区分人类和其他动物的关键能力之一。互联网上充斥着松鼠[22]和乌鸦[23]的视频，这些动物虽然可以解决一些复杂的问题以获取食物，但就解决抽象问题而言，还没有任何动物能够接近人类的水平(哪怕没有食物作为直接奖励也会去解决问题)。当然，解决问题是需要智慧的，如果我们能构建某种程序，解决人类认为难以解决的问题，这算不算人工智能走向现实的关键一步呢？因此，在黄金年代，人们在解决问题这个领域投入了大量研究，当时研究人员的标准做法是让电脑来解决人们经常在报纸趣味谜题版面能够看到的东西。我们以经典的汉诺塔为例：

在一个偏僻的寺院里，立着三根柱子，还有64个金环。每一个金环

的尺寸都不同，并且金环都可以套在柱子上。最初，所有的金环按照上小下大的顺序，都套在最左边的柱子上，然后，僧侣们要利用这三根柱子，一个一个移动金环。他们的目标是把所有金环都移动到最右边的柱子上，移动的时候，僧侣们必须遵从两个原则：

1. 一次只能把一个金环从一根柱子移动到另一根柱子上。
2. 任何一个金环都不能位于比它小的金环之上。

要解决这个问题，僧侣们必须找到正确的移动顺序，使得所有的金环从左边柱子移动到右边柱子，并且遵循大金环不能放在小金环之上的规则。

按照传说，当僧侣们完成任务的时候，世界就会毁灭[1]。当然，我们很快就会知道，哪怕这个传说是真的，也无须为此担忧，因为宇宙在僧侣们移动完所有64个金环之前早就灭亡了。所以现实中的汉诺塔只会有极少的金环，如图3所示。

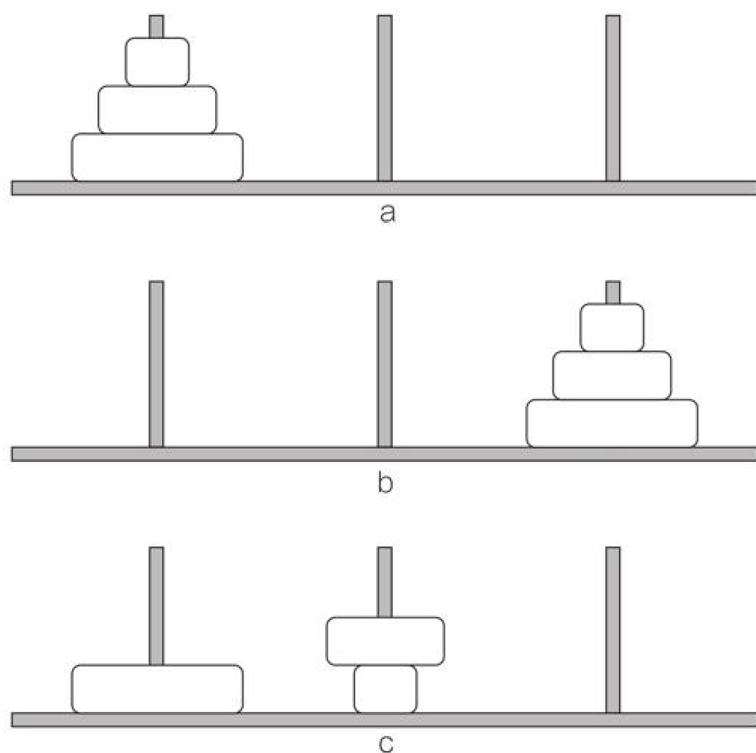


图3　汉诺塔

人工智能黄金年代的经典谜题，图a展示了汉诺塔的初始状态(所有的金环都在最左边的柱子上)，图b展示了目标状态(所有的金环都被挪到最右边的柱子上)，图c展示了错误的移动状态(大金环放在了小金环上面)。

那么，我们将如何着手解决汉诺塔这样的问题呢？答案是使用一种名叫搜索的技术。在这一点上，我得首先澄清，在人工智能领域使用“搜索”一词时，并非指打开谷歌或者百度对网站进行搜索。人工智能领域的搜索，是一种基本的问题解决技术，它涉及让系统全盘考虑所有进程。任何像下棋一样的程序都是基于搜索技术的，汽车上的卫星导航系统也是。它一次又一次地出现在诸多领域，是人工智能技术的基石之一。

所有类似汉诺塔的问题都有着相通的解决方案。在积木世界里，我们希望能够找到相应的步骤，把初始状态的物体转变为指定的目标状态。“状态”这个术语在人工智能领域指的是某个事物或者问题在某个特定时刻呈现的结构。

用搜索的方式来解决类似汉诺塔的问题，我们可以遵照以下步骤：

·首先，从初始状态开始，我们考虑每个可能的动作对初始状态的影响，执行每一步操作都会将初始状态转换为一个新的状态。

·如果其中某个新状态是我们的目标状态，那就意味着操作成功：这个问题的解决步骤就是从初始状态达到目标状态所执行的操作。

·否则，在每一个新状态上，继续考虑所有可能导致状态变更的操作，重复这样的过程，以此类推。

应用搜索的方法会得到一个树状的结构图，被称为搜索树，图4为我们展示了解决汉诺塔问题的搜索树的一小部分片段。

·最初，我们只能选择移动最小的金环，只有将它移动到中间或者最右边的柱子上。所以对应第一步，我们只有两种移动可能，以及两种不同的新状态。

·如果我们选择将最小的金环移动到中间柱子(初始状态左下方箭头对应的状态)，那么我们随后可以执行三种操作：将最小的金环移动到最左侧或者最右侧的柱子，或者将第二个金环移动到最右侧柱子(不能将第二个金环移动到中间柱子，因为它会位于最小金环的上方，这就违背了移动规则)。

·以此类推。

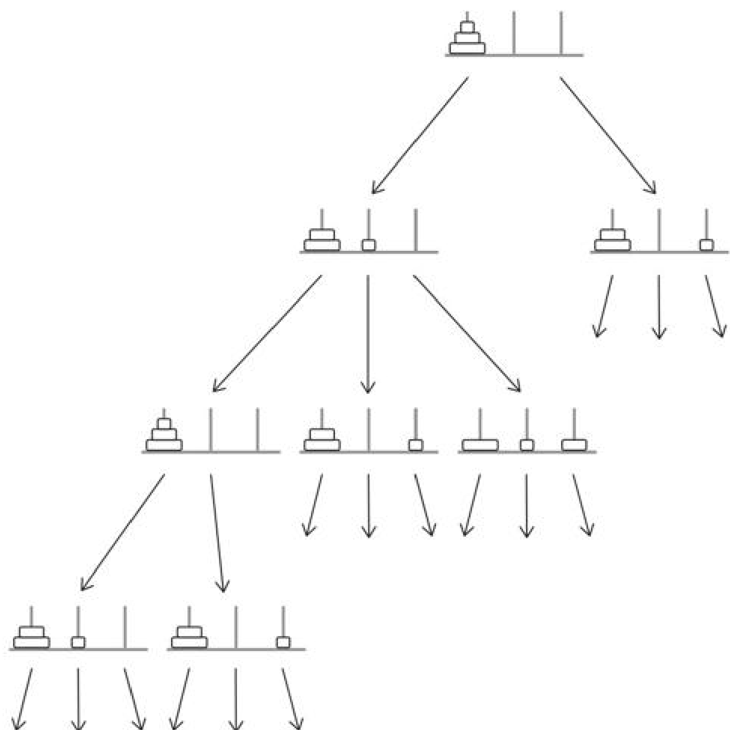


图4　汉诺塔问题搜索树的一小部分片段

因为我们逐级逐级地在生成搜索树，这就意味着我们考虑到了所有的移动可能性，因此，如果这个问题确定是有解决方案的，我们就可以确保使用搜索的方法最终能够找到它(“最终”这个词在句中是至关重要的，至于为什么，很快就会揭晓)。

那么，解决汉诺塔问题需要移动的最少次数是多少呢？也就是说，从初始状态到目标状态，最少需要移动多少次？对于三个金环而言，答案很简单，是7次。你不可能找到少于7次的解决方案了。当然，也有很多方案可以用超过七步的方式解决问题(实际上是有无穷多方案)，但它们不是最优的，因为我们想用最少的步骤、在最短的时间内解决问题。现在，考虑到我们所描述的搜索过程是详细无遗漏的——考虑到了每一步搜索树的所有可能性——搜索技术不光能保证寻找到解决方案，也能保证寻找到最优解决方案。

这就意味着，穷举搜索的方式不仅能保证在问题可解的时候寻找到问题的解，还能保证寻找到问题的最优解。另外，作为计算机的算法，穷举搜索非常简单——编写一个程序来实现它非常容易。

不过，哪怕我们只是粗略研究图4所示的片段都能够看出来，如刚才所描述的那样，采用穷举搜索的方式来解决问题是个相当愚蠢的过程。比如，仔细看看搜索树最左边的分支，仅仅移动两步以后，问题又返回到了初始状态，这是在无谓地浪费精力。如果是你来研究解决汉诺塔问题，可能会在寻找解决方案的时候犯一两次类似的错误，不过你会很快发现问题所在，并且学会避免做一些徒劳的移动。然而，一台简单执行穷举搜索的计算机却无法做到：它只能穷举出所有移动下的所有新状态，哪怕某些穷举步骤就是在浪费时间，它也会走回头路，回到已经被确认过失败的状态。

除了太多重复和低效率以外，穷举搜索还有一个根本的问题。如果你做过一些实验，会发现在大部分状态下，汉诺塔问题都有三种移动的可能性，即分支因子为3。不同的问题对应着不同的分支因子。比如，围棋的分支因子为250，这就意味着在游戏给定的任意状态下，每个玩家约有250种落子的可能性，当然，这只是平均值。所以，我们来看看搜索树随着分支因子能增长到多大——对确定了分支因子数的搜索树，在不同层级有多少种状态。以围棋为例[24]：

·搜索树的第一级有250种状态，因为在游戏棋盘的初始状态下，可以有250种新出现的状态。

·第二级搜索树有250×250 ＝ 62 500种状态，因为我们必须考虑到250种落子选择后，每一种状态都有250种可能出现的新状态。

·第三级搜索树有62 500×250，大约有1560多万种状态。

·以此类推，第四级搜索树就会有大约39亿种状态。

到此为止吧，在我撰写这篇文章的时候，一台普通的台式机已经没有足够的内存来存储四级围棋搜索树的状态了。而通常一局围棋游戏要走大概200步，在围棋搜索树中存储200步走棋的搜索树状态数目，那是一个大到你我都无法想象的天文数字，它比我们宇宙中所有原子的数目还大几百个数量级。无论对传统计算机技术做出怎样的改进，它们都无法完成这样可怕的搜索树。

搜索树这种迅速、荒唐的增长速度导致了无法想象的问题，我们称之为组合爆炸，它是人工智能所面临的最重要的实际问题，因为搜索技

术在人工智能领域使用得非常广泛[25]。如果你能够寻找到一个快速并且万无一失的办法解决搜索树的难题——发明一种方法，既能达成穷举搜索的目标，又无须占用这么恐怖的资源——恭喜你，你将青史留名，并且让许多人工智能目前所面临的困难迎刃而解。可我不得不遗憾地说，你做不到。我们没办法回避组合爆炸的问题，只能想办法处理它。

在人工智能发展初期，组合爆炸被认为是根本性问题，麦卡锡将其确定为1956年人工智能暑期学校的重要研究课题之一。人们的关注点主要集中在提高搜索效率上，对此，有几种不同的解决方案。

一种是以某种方式集中搜索。其中一种典型的方式是并非逐级建立完整的搜索树，而是沿着其中一个分支构建搜索树。这种方式被称为深度优先搜索。通过深度优先搜索，我们沿着一个分支往下扩展，直到得到解决方案或者确信无法得到解决方案。如果遇到困难(比如类似图4中最左边的分支，又回到一个已经出现过的状态)，那么我们就停止在该分支上的扩展，返回到上一级，选择另外的分支。

深度搜索的主要优点在于不用存储整个搜索树，只需要存储当前正在处理的分支即可。但它有一个很大的缺陷：如果选择了错误的分支进行探索，可能会在错误的路上越走越远，永远找不到解决方案。所以，要想使用深度优先搜索，首先我们得确认哪个分支最值得搜索。这时候，我们就需要启发式搜索来帮忙了。

启发式搜索的概念就是使用所谓的“经验法则”来指导搜索的重点。通常我们也无法寻找到直指正确搜索路径的启发式方法，但我们往往可以在感兴趣的问题上找到启发式搜索方向。当然，我们也明白，有些情况下，启发式搜索的运行情况并不那么尽如人意。

启发式搜索是一个自然而然产生的概念，多年来它被重新定义了很多次，所以争论到底是谁发明了它已经毫无意义。但是，我们可以确定，第一个应用在人工智能程序的启发式搜索来自一个玩跳棋游戏的程序[26]，由IBM的亚瑟·塞缪尔(Aithur Samuel)于20世纪50年代中期创造。塞缪尔编写的跳棋程序，在许多方面为人工智能领域开辟了新天地。首先，以棋盘为人工智能技术的试验场，这一传统就是塞缪尔的跳棋程序开创的。即使在今天，这也是一种常见的技巧，虽然也有不少人提出反对意见。其次，正如我们所提到的，也是即将展开详细讨论的一点，该程序是基于启发式搜索编写的。最后，我们会在之后详细讨论，

塞缪尔的跳棋程序是第一个真正意义上的机器学习程序：他的程序自学了怎么玩跳棋。

塞缪尔跳棋程序的关键点在于给棋盘的各个位置赋予不同的权重，用以评估对选手而言位置“好”的程度：从直觉来说，某些“好”的位置可能让某位选手更容易取得胜利，而一些“坏”的位置则会导致失败。为此，塞缪尔用一系列特征值来计算棋盘上的点位价值。例如，某些比较关键的位置，能够让你形成连跳，这样的连跳越多，就意味着你在棋盘上的形势越占上风。然后，程序将不同的评估参数整合起来，给出棋盘位置的一个综合评估价值分数，形成一个量化的评估标准。再根据启发式搜索方法来选择最佳的棋盘移动路径。

在实践中，要考虑比简单的启发式算法更复杂的情况：因为跳棋是一种对抗游戏，你必须考虑到对手可能如何行动。塞缪尔的跳棋程序会认为你的对手会做出对你最不利的举动。这种“最坏情况推理”的方法(即假设对手的行为会使得自己的得分最大化，让你的得分最小化)称为极大极小值搜索，是对抗性游戏的基本概念。

塞缪尔的程序玩跳棋的水准相当高，从许多方面来看都是一个令人惊叹的成就。尤其是考虑到当年所使用的计算机配置——塞缪尔使用的是IBM701计算机，只能处理相当于几页文本长度的程序代码——那就更了不得了。现代的台式计算机内存是当年的数百万倍，运行速度也是数百万倍。在当年计算机条件如此有限的情况下，跳棋程序能够达到这样的水准，简直是超乎想象。

大部分早期的启发式搜索算法，包括塞缪尔的程序，都使用了某种特别的启发方式——“尝试—实践—判断”。20世纪60年代末，美国SRI人工智能研究中心的尼尔斯·尼尔森(Nils Nilsson)和他的同事取得了突破，他们开发出名为A^*的技术，作为我们之前讨论过的SHAKEY项目的一部分。A^*定义了一些简单的规则，让我们可以确定哪些启发是“有益的”。在A^*之前，启发式搜索不过是一个猜谜游戏；而在A^*之后，它是一个数学上很容易理解的过程[27]。A^*现在被视为计算机科学中的基础算法之一，并且在实践中得到了广泛的应用。其实，我们在生活中经常会遇见使用A^*算法的程序，比如车载卫星导航系统等。尽管A^*很惊艳，但它仍然依赖于使用特定的启发式搜索：好的启发能很快地找到解决方案，而差的启发就没什么价值。对如何为特定问题找到更优秀的启发，A^*本身并没有给出答案。

人工智能遇上了复杂性障碍

我们在上一章讲过，计算机的出现源自艾伦·图灵想解决一个彼时数学界的世纪难题。图灵发明计算机可能算是计算机科学史上最大的讽刺之一，因为他的本意是证明有些事情计算机永远也办不到。

在图灵的研究成果问世后的几十年里，探索计算机能做和不能做的事情的范围形成世界各地数学系的小分支的研究方向。这项工作的重点在于把那些固有的不可判定问题(无法用计算机解决的问题)和可判定问题(能够用计算机解决的问题)区分开来。从那时起，人们发现了十分有意思的现象——不可判定的问题存在层次结构：即存在某些问题，不仅不可判定，还是高阶不可判定的(这对某些研究特定领域的数学家而言很棘手)。

不过在20世纪60年代，确定一个问题是否可判定，还远远不够。一个问题在图灵看来是可以解决的，并不意味着它在实际操作上是可以实现的：图灵只是从理论上证实某些问题有解，但是有些问题的解决方式需要消耗庞大的计算资源——需要动用的计算机内存大到不可估量，或者计算机的运行速度慢得出奇，能计算出结果的时间遥遥无期。很明显，许多人工智能问题陷入了这样尴尬的境地。

下面就是一个例子，按照图灵的理论，很容易证明它能被解决：

> 你的办公室里有四个才华横溢的人为你工作：约翰、保罗、乔治和林戈。现在有个项目，你需要成立一个三人团队。但很遗憾，由于个人竞争因素，约翰和保罗不能跟彼此合作。
>
> 你能组建一支合适的队伍吗？

这道问题的答案显然是“能”，你可以选择约翰、乔治和林戈，或者保罗、乔治和林戈(请注意，这是该问题仅有的两种解法)。

然后我们再加上一个限制：约翰和乔治不能一起工作。那么答案就明显了：保罗、乔治和林戈。

现在我们加上最后一个限制：保罗和乔治不能一起工作。那么我们的答案就变为“不能”，因为没有一种组合可以同时满足三个“禁止组队”的要求。

我们用抽象的方式来描述这个问题[28]：

首先给出一个包含n个人的列表(在上述例子中，这些人就是约翰、保罗、乔治和林戈，因此n ＝ 4)，以及一个“禁止组队”的列表(例如，“约翰和保罗不能在一起工作”)。那么，我们能否得到一个包含m个人的团队，使得所有禁止组队的条件都能得到满足？(显然，要使得这个问题有意义，m必须小于n。)

必须指出的是，这个问题在原则上是很容易解决的。简单地列出所有含有m个成员的团队名单，并检查每一个团队名单是否与禁止组队列表有冲突即可。用计算机编程来做这个步骤非常容易。因此，在图灵的观念里，这个问题很容易求解。

那么，问题的关键来了，我们需要检查多少种组合的可能性呢？如果是4人团队里面选3人，那很容易，你需要检查4个有可能的选择。但当团队人数增多，会出现什么情况，我们来看看吧。

如果是10人团队选5人，你就得检测252种可能性。好吧，很枯燥，但是还能忍受。不过可以看到的是，选择可能性的增长速度远远大于整个团队成员或者目标团队成员的数量增长的速度。

如果是100人团队里面选50人，那你就得检测大约10^{29}这么多种可能性了。一台当代的大型计算机每秒可以评估大约10^{10}种可能性——听起来挺多的，但是，稍微计算一下你就能意识到，即便算到宇宙末日，也无法检测完毕。期待英特尔的工程师们提供计算速度更快的芯片也无济于事：传统计算机技术再怎么突飞猛进，也无法在合理的时间内完成这个数量级的计算任务。

我们在这里所看到的现象也是组合爆炸的例子。在研究搜索树的时候我介绍过组合爆炸的概念，搜索树中的每个层级都呈指数型增长。所以在层级增多的时候，组合爆炸会导致可能出现结果的增长速度超乎想象。在团队建设的案例中，只要团队总人数增加1个人，我们必须考虑的潜在组合型就会翻倍。

所以，这种彻底列举每一种组合可能性的方式，是不可行的。理论上它行得通(如果时间足够，总会得到正确的答案)，但在实践中它毫无意义(因为对每种可能的组合做判断需要的时间是天文数字)。

正如我们之前提到过的，简单的穷尽式搜索是一项非常原始的技术。我们可以使用启发式方法来对它进行优化，但是并不保证一定有效。有没有一种更聪明的方法可以寻找到合适的组合，又不用彻底检查所有的备选方案？很遗憾，答案是没有。你会发现对搜索方式的优化可

以改善一些东西，但是，最终你仍然无法绕过组合爆炸。在大多数情况下，能在理论上确保问题可以解决的方案，在实践中都会遇到这个问题。

因为我们的问题是一个**NP**完全问题的案例[2]。恐怕这个首字母缩写对不熟悉计算机编程的人而言太过陌生："NP"代表"不确定多项式时间"，具体的技术定义相当复杂。幸运的是，NP完全问题背后的逻辑很简单。

组合问题是一个NP完全问题，就像我们团队建设的例子。在这个问题中你很难找到解决方案(我们在上文讨论过，因为要彻底检查每一种组合需要耗费无尽的时间)，但是又很容易验证是否找到了解决方案(在这个示例中，我们可以简单地验证它是否与禁止组队的规则相冲突)。

NP完全问题还有一个重要特征，为了理解它，我们要引入另一个问题(请相信我，这个真的很有必要)，这个问题相当著名，你可能听说过它。它被称为旅行推销员问题[29]：

> 有一位推销员要去被道路连接的若干城市推销商品，走完所有城市以后他必须回到起点。并非所有城市都有公路直接相连接，而有公路连接的城市，推销员知道每条公路的具体长短。请问：推销员走遍所有的城市，最终回到起点，怎样选择一条最短的路径？

这个问题跟团队建设类似，都是组合类问题。我们可以用蛮力计算来解决它，只要列出所有可能的路线，并检查每条路线的长短，然后选择最短的一条即可。然而，正如你现在所想的那样，随着城市数量的增加，可能存在的行进路线会呈指数级增长：如果要走遍10座城市，就得考虑360多万种可能性，11座城市就会激增为4000万种。

所以，旅行推销员问题和团队建设问题一样，都会面临组合爆炸的难题。但除此之外，这两个问题似乎没什么共通之处，毕竟团队建设和寻找最短路径有什么关联呢？但是在NP完全问题的范畴内讨论的话，这两个问题，虽然看上去差别很大，但本质上是同一个问题。

我这句话的意思可以这么理解：假设我已经发明了一种巧妙的方法，可以保障高效又准确地找到团队建设问题的正确答案，现在你给我一个旅行推销员问题的案例，我就可以把这个旅行推销员问题转化为团队建设问题的一个特例，我的方法可以迅速解决它，你则可以得到想要的答案。这就意味着，如果你发明了一种可以高效又准确解决团队建设

问题的方法，那么就等同于你发明了可以高效又准确解决旅行推销员问题的方法。这种神奇的方法并不仅仅适用于这两个问题，而是可以解决所有NP完全问题。

所有的NP完全问题都能共通，可以互相转换。这就意味着一个很明显的事实：要么所有NP完全问题存在通用的解，要么它们都无解。如果你能找到高效又准确解决某个NP完全问题的方法，就意味着你能够解决所有NP完全问题。但是，到目前为止，还没有任何人寻找到解决NP完全问题的高效准确的方法。事实上NP完全问题能否有效解决，是当今科学界最重要的开放性问题之一。这个问题被称为**P**与**NP**问题[30]，如果你能证明这个问题是否有解，并且能让科学界认同，你将获得克雷数学研究所提供的100万美元奖金，正是该研究所于2000年将其确定为“千年问题”之一。精明的投资商认为NP完全问题不能得到有效解决，同样也是精明的投资商告诉我们，要确切知道这一点，还有很长的路得走。

如果你发现自己要解决的问题是NP完全问题，这就意味着传统意义上的计算机技术在解决该问题上是行不通的：从精准的数学意义来说，你的问题太难了。

NP完全问题的基本结构早在20世纪70年代就已成形。1971年美籍加拿大数学家斯蒂芬·库克(Stephen Cook)的一篇论文确定了NP完全问题的核心思想，随后美国人理查德·卡普(Richard Karp)证明了库克NP完全问题的范畴比最初想象的要广泛得多。整个20世纪70年代，人工智能领域的研究人员开始利用NP问题完全性理论来研究他们的课题，结果令人震惊。不管哪个领域——解决问题、玩游戏、计划、学习、推理任何方面——似乎关键性的问题都是NP完全问题(甚至更糟)。这种现象普遍得成了一个笑话——所谓的“AI完全问题”[3]意味着“一个和AI本身一样难的问题”，如果你能解决一个AI完全问题，你就能解决所有AI问题了。

发现你正在研究的问题是NP完全问题——或者更糟——真是个沉重的打击，在NP完全性理论及其论证结果被理解之前，人们总是希望出现重大突破使得这些问题变得简单——用术语来说就是易处理。从技术上讲，这种希望还是存在的，因为我们还没有明确地证明NP完全问题不可解。但到了20世纪70年代，NP完全性理论和组合爆炸成为笼罩在人工智能领域的阴影，以计算复杂性的形式阻拦在人工智能发展道路上，使它陷入停滞。黄金年代发展起来的技术似乎都无法扩大它们的使用范围，走不出玩具或者特定小世界(比如积木世界)的范畴。50到60年代过

于乐观的预测遇到了现实难题，这让人工智能领域的先驱者们无比困扰。

人工智能等于炼金术？

20世纪70年代初，人工智能的瓶颈让科学界越来越沮丧，没有太多有用的核心研究进展，而某些研究人员又大肆鼓吹人工智能的未来，于是，到了70年代中期，批评人工智能的风潮达到狂热的程度。

在所有批评者中——相信我，这个数量非常大——最直言不讳、出语刻薄的无疑是美国哲学家休伯特·德莱弗斯(Hubert Derytus)。在20世纪60年代中期，他受兰德公司委托撰写了一份关于人工智能研究进展状况的报告，德莱弗斯选择的报告题目是《炼金术与人工智能》，明白无误地表达了他对这一领域及其工作人员的蔑视。公开地把一门严肃的科学和炼金术类比，这是非常侮辱人的。现在，重读《炼金术与人工智能》，哪怕时隔半个世纪，我仍然不得不说从来没见过这样的报告，其内蕴的蔑视之意如此明显，令人震惊。

虽然我们不赞成德莱弗斯批判人工智能的方式，但很难回避这个现实：他的某些观点是有道理的，特别是关于人工智能先驱者们浮夸的观点和宏伟的预测。赫伯特·西蒙(后来获得了诺贝尔经济学奖)在1958年写道：

> 这并非耸人听闻，但是……现在世界上出现了能够思考、学习和创造的机器。在可以预见的未来，它们的能力将迅速增强，能够处理的问题范围将迅速扩大到人类思维的范围。

在当时，这样的说法很多，我引用西蒙的观点是因为他最有名气。虽然西蒙是人工智能乐观主义者，但他绝不是最狂热和不讲理的倡导者。令人痛心的是，事后看来，他的乐观预测是不切实际的。时至今日，人工智能界的一些成员认为，当时的研究人员之所以做出这样的乐观声明仅仅是因为他们被兴奋冲昏了头脑——也有人持更为愤世嫉俗的观点，称这种大肆炒作的目的是吸引投资和研究基金。

不管怎么说，我的专业敏感性让我必须声明一下，虽然当年的研究人员乐观得近乎天真，但并没有出现大规模蓄意欺骗或者歪曲人工智能研究状态的事情。研究人员之所以乐观过头，是因为他们热爱这门学科，并且真的相信——也希望其他人相信——他们确确实实建立了可以学习、计划和推理的程序，尽管有诸多局限性。他们只是乐观地认为，

扩大这些程序的应用范围应该很容易。

所以，我认为某些乐观过头的情绪是可以被谅解的。尤其是，当年的研究人员没有意识到正在研究的问题本质上是计算机无法处理的，因而总存在出现突破性进展，使得问题迎刃而解的可能性。但当NP完全问题无解的概念开始普及以后，社会各界才慢慢明白，人工智能所面临的问题有多艰难。

黄金年代的终结

1972年，英国一个重要的科研资助机构要求著名数学家詹姆斯·莱特希尔(James Lighthill)爵士评估人工智能的现状和前景。据说因当年世界最先进的人工智能研究中心之一(如今也是)——爱丁堡大学人工智能社区出现学术斗争的相关报道，所以才要求做这样的评估。莱特希尔荣列当年剑桥大学卢卡斯数学教授席位，这是英国数学家所能担任的最负盛名的学术职务，斯蒂芬·霍金(Stephen Hawking)后来也曾担任过。莱特希尔在实际应用数学方面有着丰富经验，时至今日，重读莱特希尔报告，可以看出他为当年人工智能的繁荣感到相当迷惑：

> 那些能力卓绝又受人尊敬的科学家写道：……所谓的人工智能……代表了进化过程中的另一条路。20世纪80年代，在人类规模的知识基础上，会实现多用途的人工智能。预测结果令人惊叹：到了2000年，机器智能将全面超越人类。

莱特希尔在报告中对当时的主流人工智能相当不屑一顾，他明确指出组合爆炸是人工智能领域无法解决的关键问题之一。他的报告立刻导致英国各地对人工智能研究经费的严重削减。

在美国，人工智能研究的资金大多来自军方，以国防高级研究计划局(DARPA)及其前身高级研究计划局(ARPA)为主。但到了20世纪70年代初，因为未能兑现诸多乐观的承诺，美国对人工智能研究的资助也随之削减。

20世纪70年代初到80年代初这10年被称为人工智能寒冬，更确切地说应该是人工智能的第一个寒冬，因为此后还有类似事件发生。人们得出了一个普遍结论，人工智能研究人员对这一领域充满了盲目乐观和毫无根据的预测，结果什么都没有。当时人工智能的名声就像医学界的顺势疗法[4]，作为一门严肃的学科，它似乎正在走向衰落。

第三章 知识就是力量

我们不能低估20世纪70年代中期人工智能发展遇到的重挫，许多学者将人工智能视为一门伪科学——直到最近这一领域才算是恢复了名誉。不过，就在莱特希尔报告广为流传并且造成严重后果之时，一种新的研究方法开始引起人们的注意，它宣称可以解决人工智能受人诟病的问题。20世纪70年代末到80年代初，不少研究人员都突然决定朝这个方向进行研究，他们认为人工智能领域此前存在的问题是过度关注搜索和解决问题这种通用法则。有人提出，这些“弱”方法缺少一个关键的要素，而这一要素才是在所有智能行为中起决定性作用的组成部分：知识。Scientia potentia est(知识就是力量)，17世纪哲学家弗朗西斯·培根(Francis Bacon)如是说。人工智能研究者从字面上理解培根的名言，他们支持以知识为基础的人工智能研究，确信捕捉和使用人类现有的知识才是人工智能进步的关键。

一种基于知识的人工智能系统——专家系统开始出现，它能利用人类专业的知识来解决特定的、狭义领域的问题。专家系统提供的证据证明，人工智能在完成某些特定领域的任务方面远胜人类，更重要的是，它们首次向人们证明，人工智能可以应用于商业领域。基于知识的人工智能系统可以向广大受众传授相关的技术，这一代的人工智能研究毕业生决心把他们的知识应用在此领域。

专家系统与通用人工智能不同，它的目标是解决非常狭义、非常具体的问题，解决这类问题通常需要相当专业的知识。通常，能够解决这类专业问题的人类专家都需要花费极长的时间来学习相关知识，而这类专家相当稀少。

在接下来的10年里，基于知识的专家系统是人工智能研究的主要焦点，工业界的巨额投资流入了这一领域。20世纪80年代开始，人工智能的冬天趋于结束，另一股更狂热的人工智能浪潮悄然到来。

在本章中，我们将了解从20世纪70年代末期到80年代末期蓬勃发展的专家系统，我会从如何获取人类专家的知识并将其输入计算机开始，为你讲述MYCIN[5]的故事——它是当年最著名的专家系统之一。我们将看到研究人员如何利用数学逻辑的强大和精准性，试图建立更丰富的

获取知识的方法，以及为何这个目标最终也落空了。接下来，我们将听到有关**Cyc**工程的故事，这是人工智能历史上最雄心勃勃，也是最臭名昭著的失败项目之一。它试图利用人类专家的知识来解决有史以来最大的难题：通用人工智能。

使用规则获取人类专家知识

在人工智能系统中加入知识不是什么新鲜想法，正如我们在前一章所看到的，启发式方法作为一种将解决问题的重点放在有希望的方向上的方法，在黄金年代被广泛应用。启发式方法可以理解为包含能够解决问题的知识，但是启发式方法并没有直接去获取知识。而基于知识的人工智能体系拥有一个全新的思想：人工智能系统应该明确地获取和展示人类解决某类问题的专业知识。

最常见的方案是基于规则的，被称为知识表述。人工智能环境下，一条规则以“如果……那么……”的形式获取离散的知识块。实际上规则相当简单，我们来举个例子说明。以下是一些规则(用自然文字编写，而不是代码)，它们是人工智能民间传说的一部分[31]。下面是为了帮助人们对动物进行分类而设计的人工智能专家系统，看看它是如何获取知识的(有关如何应用规则的细节，请参见附录A)：

> 如果该动物产奶，那么它是哺乳动物。
>
> 如果该动物有羽毛，那么它是鸟类。
>
> 如果该动物能飞并且能产卵，那么它是鸟类。
>
> 如果该动物吃肉，那么它是肉食动物。

我们对这些规则作出如下解释：每一条规则都有前因条件(即“如果”之后的部分)和后果结论(即“那么”之后的部分)。例如以下规则：

> 如果该动物能飞并且能产卵，那么它是鸟类。

该规则中的条件是“该动物能飞，该动物能产卵”，结论就是“它是鸟类”。如果我们当前所掌握的信息与条件项相匹配，那么规则就会被触发，我们就能根据这条规则得出结论。因此，要使这条规则生效，需要两个条件：我们试图分类的那个动物能飞，而且能产卵。如果我们确实掌握了这两条信息，那么规则成立，我们就能得出结论，该动物是鸟类。这个结论为我们提供了更多的信息，这些信息又可以用在后续的规则中，用以获取更多的信息，如此层层推进。通常，专家系统以咨询的

形式与用户交互，用户负责向系统提供信息，并且回答系统提出的问题。

图5所示的就是一个典型的专家系统结构，其中，知识库包含系统所拥有的知识——那些规则。工作存储器则包含了系统拥有的，有关当前正在解决的问题信息(例如“该动物有毛发”)。最后，推理机则是专家系统的一个重要组成部分，它负责在解决问题的时候应用系统内存储的知识。

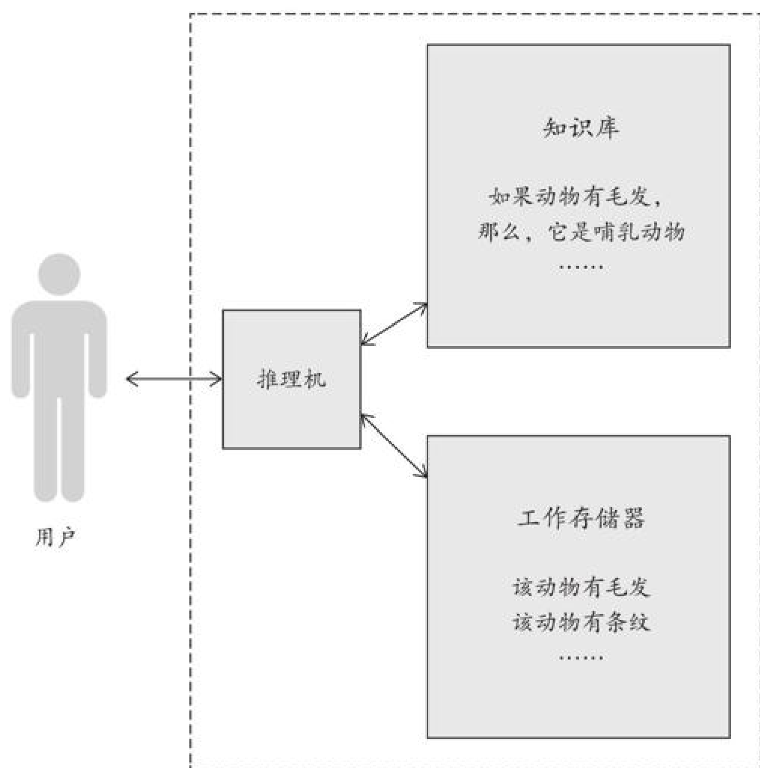


图5　经典专家系统结构

只要给定知识库，比如上文举例的动物分类知识，推理机就能够以两种方式运行：第一种，用户向系统提供他们所知道的有关问题的信息(以动物分类为例，用户向系统描述该动物是否有条纹、是否食肉等)，推理机会根据用户提供的信息，应用规则去获取尽可能多的新信息，这

个过程叫作规则触发。然后，推理机将触发规则以后获得的新信息添加到工作存储器中，继续查看是否有新的规则被触发，然后不断重复这个过程，直到彻底无法通过已知信息应用更多规则得出更多新信息为止。这种方法被称为正向推理：从数据推理到结论。

另一种方式是反向推理，即我们从想建立的结论开始，反向推理出数据。例如，我们想确定这个需要被分类的动物是否为肉食动物，规则告诉我们，如果动物吃肉，就是肉食动物，因此，我们可以尝试确定动物是否吃肉。因为没有任何规则的结论是动物是否吃肉，系统只能向用户直接询问这个问题。

MYCIN：一个经典的专家系统

在20世纪70年代出现的第一代专家系统中，最具代表性的可能就是MYCIN系统了(这个系统名称来源于抗生素的英文词后缀“mycin”)[32]。MYCIN系统首次证明，人工智能在某些重要的领域表现可以优于人类专家，它为后来无数的专家系统提供了模板。

MYCIN本来是用于辅助医疗的系统，为人类血液疾病的诊断提供专业建议。它是由斯坦福大学的一个研究小组开发的，包括布鲁斯·布坎南(Bruce Buchanan)领导的人工智能实验室专家组和特德·肖特利夫(Ted Shortliffe)领导的斯坦福医学院专家组。事实上，这个项目成功的一大要素就在于，专家系统是由真正的人类专家参与建设的。后来有许多专家系统都宣告失败，因为它们缺乏了相关领域人类专家的必要支持。

跟我们在前文提到的动物分类规则相比，MYCIN关于血液疾病的知识则是用稍微丰富一些的规则来表示的。典型的MYCIN规则(用自然语言表述)如下：

如果：

·该有机体不会被革兰氏染色法染色，并且

·该有机体的形态是杆状，并且

·该有机体是厌氧的

那么：

有可能(60%)该有机体是类杆菌

这是用实际的人工智能语言表述的MYCIN规则：

```
RULE036:
PREMISE: ($AND (SAME CNTXT GRAM GRAMNEG)
(SAME CNTXTM MORPH ROD)
(SAME CNTXT AIR ANAEROBIC))
ACTION: (CONCLUDE CNTXT IDENTITY
BACTEROIDES TALLY 0.6)
```

研究人员大概用了5年的时间对MYCIN的知识库进行编码和补充，在它的最终版本里，知识库已经包含了数百条规则。

MYCIN之所以被称为最具代表性的专家系统，是因为它包含了后来的专家系统必不可少的所有关键特性。

首先，MYCIN的系统操作和人类专家进行交互类似——向用户提出一系列问题，并且记录用户的响应。这成为专家系统的标准模型，而MYCIN的主要功能——诊断——则成为专家系统的标准任务。

其次，MYCIN的推理是可以还原和解释的。推理透明度的问题在人工智能应用方面有时极其重要。如果一个人工智能系统被应用在一些生死攸关的决策中(以MYCIN而言，比如治疗方案事关病人的生死)，要想让人们遵从它的建议，就需要人们信任这个系统。因此，解释和证明人工智能建议的合理性是非常有必要的。经验表明，作为“黑盒”运行的系统，如果没有能力证明其建议的合理性，就会受到用户的严重质疑。

MYCIN至关重要的能力是它可以明确给出得出这一结论的原因，它是通过一系列推理链来得出最终结论的，即那些被触发的规则和触发规则的信息，都是有迹可寻的。在实践中，大多数专家系统的解释能力最终都归结为类似的东西。虽然不是特别理想，但这样的解释很容易追根溯源，并且有助于理解系统运作的机制。

最后一点，MYCIN能够应对不确定性：有些时候用户向系统提供的信息并不是完全真实和准确的。应对和处理不确定性是对专家系统及人工智能系统的一个普遍要求，在MYCIN这样的系统中，极少根据某个单一的特性就得出明确结论的规则。例如，用户的血液检测结果呈阳性，这就为系统判断提供了一个有力的证据。但是，总有出现检测错误的概率(比如“假阳性”或者“假阴性”之类)；或者，患者表现出的某些症状可能是某种特定疾病的征兆，但是不能确定它一定就是某种特定疾病(例如

咳嗽是拉沙热的典型症状之一，但病人有咳嗽的症状不能够直接判定他患了拉沙热)。为了能够做出准确的判断，专家系统需要以某种更保险的方式来考虑症状信息。

为了处理不确定性，MYCIN引入了确定性因素的技术——某种表示对某一特定信息的信任或者不信任程度的数值。确定性因素技术是处理不确定性问题的一个相当特别的解决方案，但也因此招来了大量的批评。不确定性处理问题和推理问题成为人工智能研究的一个重要课题，至今仍然如此。我们将在第四章深入讨论。

1979年，MYCIN参与了10个实际病例的评估实验，在血液疾病诊断方面，MYCIN的表现与人类专家相当，并且高于普通医生的平均水平。这是人工智能系统首次在具有实际意义的任务中展示出人类专家级或以上的能力。

繁荣再起

20世纪70年代涌现的专家系统绝不止MYCIN，同样来自斯坦福大学，由爱德·费根鲍姆(Ed Feigenbaum)领衔的项目DENDRAL才是世界上第一例成功的专家系统，它也使费根鲍姆成为知识型系统最著名的倡导者之一，并被人们尊称为“专家系统之父”。DENDRAL的开发目的是帮助化学家根据质谱仪提供的信息来确定化合物的成分及结构。在20世纪80年代中期，每天有成百上千的人在使用DENDRAL。

美国数字设备公司(DEC)开发的R1/XCON系统，旨在帮助他们配置VAX系列计算机。在20世纪80年代，DEC公司声称该系统已经处理了超过8万份订单，当时系统有3000多条规则，涉及5500个不同的系统组件。到了80年代末，R1/XCON系统的规则数量已经扩展至17 500多条，开发者声称该系统为公司节省了超过4000万美元。

DENDRAL项目证明专家系统是可用的，MYCIN证明它可以在专业领域胜过人类专家，R1/XCON证明了它有商用价值。这些成功的故事引起了许多人的兴趣：人工智能似乎已经为商业化做好了准备。这是一个激动人心的故事，毫无疑问，海量的研发资金纷纷流入这个领域。一批初创公司忙不迭地从再度繁荣中获利，典型案例就是用于构建专家系统的软件平台，以及开发和部署专家系统的支持服务。不过你也可以购买专门设计用来快速执行LISP——构建专家系统的首选编程语言——的计

算机，这些LISP专用机器一直到90年代早期还在使用。后来，普通个人计算机价格日益便宜，功能也日益强大，那时候再花费7万美元买一台只能运行LISP的计算机就是毫无意义的了。

不仅仅是软件行业，其他行业也争先恐后地乘着这场人工智能再度繁荣的东风飞速发展。20世纪80年代，工业界开始意识到，知识体系，尤其是专业知识，是可以培育和发展的重要资产，可以带来高额利润。专家系统似乎使这种无形的资产变为有形，知识体系的概念与当时西方经济进入后工业时代的观点相呼应——在这一时期，经济发展的新机遇主要来自所谓的知识型产业和服务业，而不是制造业或其他传统产业。

回顾一下专家系统发展的经历，我们可以看到，人工智能的再次繁荣不仅仅是MYCIN、DENDRAL等案例成功的故事，更重要的是，专家系统赋予了人工智能另一种可能性。你不需要相关知识领域的博士学位，就能构建一个专家系统(我敢打赌，你可以轻而易举理解前文所述的动物分类规则)。任何熟悉编程的人都可以明白专家系统的原理，事实上，构建一个专家系统似乎比传统编程还容易一些。一个全新的职业出现了：知识工程师。

具有讽刺意味的是，1983年，英国政府发起了一场雄心勃勃的计算机技术研究资助计划，名叫“阿尔维计划”，计划的核心就是发展人工智能。但鉴于莱特希尔10年前的报告几乎让英国的人工智能产业陷入死局，考虑到该领域的负面名声，似乎所有参与阿尔维计划的人都不愿意将之称为人工智能。相反，他们称之为“基于知识的智能系统”。人工智能的未来似乎一片光明——只要你不把它叫作人工智能。

基于逻辑的人工智能

虽然规则成为专家系统获取人类知识的主要方法，但也有大量其他方案存在。例如，图6就展示了一种名为脚本的知识展示方案(简化)版本，由心理学家罗杰·尚克(Roger Schank)和罗伯特·亚伯森(Robert P. Abelson)开发。该方案基于一种关于人类理解能力的心理学理论建立，理论指出，我们的行为部分受刻板印象模式(即“脚本”)支配，我们也用这些模式来理解世界。他们认为同样的模式可以应用在人工智能中。以图6所示脚本为例：这是一个典型的餐厅场景，脚本(用自然语言)描述了参与者的各种角色(顾客、服务员、厨师、收银员)、启动脚本所需要的条件(客户饿了)、脚本需要操作的各种物理项(食物、桌子、钱、菜单、

小费)，并且，最重要的是与脚本相关的常规事件序列，在图6中，这些事项的编号为1—10。尚克和亚伯森推测这样的脚本可以用在人工智能程序中，用于理解故事。他们认为，当故事中的事件与常规剧本不同的时候，故事就会变得有意思(有趣、可怕、令人惊讶)。例如，一个有意思的故事可能包含这样的情况：在第4步以后脚本就停止运行了，即顾客点了餐，但是没有上菜。一个违法故事可能省略第9步：顾客点餐并吃东西，但没有付钱就离开了餐馆。他们尝试了几次以脚本为基础构建能够理解故事的系统，但成功率有限[33]。

脚本名称：餐馆

角色：顾客、服务员、厨师、收银员

启动条件：顾客饿了

道具：食物、桌子、钱、菜单、小费

事件：

1. 顾客进入餐馆
2. 顾客找了张桌子坐下来
3. 服务员拿菜单
4. 顾客点餐
5. 服务员上菜
6. 顾客吃东西
7. 顾客向服务员要账单
8. 服务员把账单给顾客
9. 顾客把钱付给收银员
10. 顾客离开餐厅

主要事件：6

结果：顾客吃饱了

顾客的钱减少了

餐馆的钱增加了

图6　描述典型的餐馆就餐经历的脚本

另一个引起广泛关注的方案是语义网[34]，它非常直观、自然，在当今社会，也经常被重新定义——事实上，如果让你去发明一个知识表述方案，我认为你很有可能做出类似的产品。图7展示了一个简单的语义网，它代表了有关我的一些知识(我的出生日期、居住地、性别和下一代)以及一些世界性常识(比如：女人首先是人，大教堂既是一座建筑又是一个礼拜场所等)。

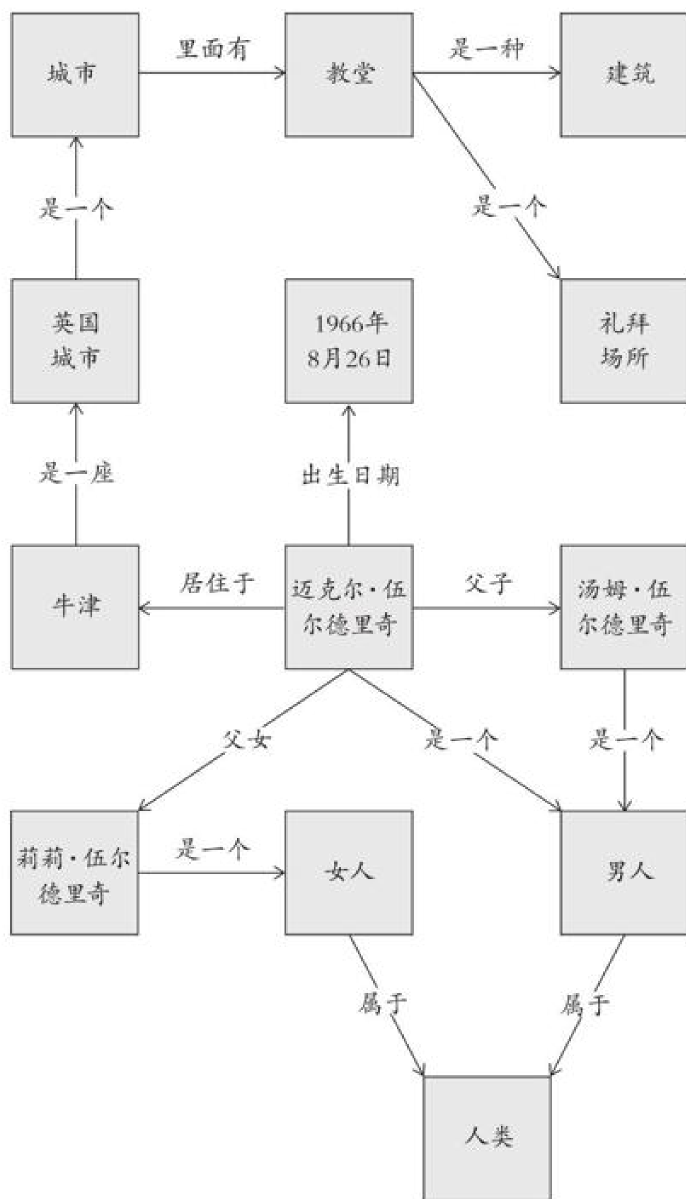


图7　一个简单的语义网，包含我的个人信息、子女信息和居住信息

在以知识为基础的人工智能兴起时，似乎每一个人都有自己的知识表述方案，而且跟其他人的不兼容。

虽然专家系统事实上的知识表述方式是基于规则的，但在知识表述方面，研究人员仍然有不少困扰。其中之一就是规则太简单，无法获取复杂环境下的相关知识。比如MYCIN系统的规则就不适用于会随时间变化的环境，也不适用于多个用户(不管是人类还是人工智能)的环境，或者实际状态存在各种不确定性的环境。另一个问题是，用于专家系统获取知识的各种方案似乎都有些武断，研究人员希望能了解专家系统中的知识实际上意味着什么，并确保系统进行的推理是可靠的。简而言之，他们想为基于知识的专家系统提供合适的数学基础。人工智能研究人员德鲁·麦克德莫特(Drew McDermott)在1978年撰写的一篇文章中总结了这些问题，并称之为“无意指不表达”[35]，“对一个系统而言，正确性固然是重中之重，”他写道，“让人能理解它，也是十分关键的。”

20世纪70年代末就开始出现相关的解决方案，即使用逻辑作为知识表述的统一方案。为了理解逻辑在基于知识的系统中所扮演的角色，我们有必要了解一些逻辑的知识，以及它如何作用于系统。逻辑学的发展是为了理解推理，特别是区分好的(完备的)推理和坏的(不完备的)推理。让我们看一些推理方面的例子，包括完备和不完备的。

凡人终有一死；

艾玛是凡人；

所以，艾玛终有一死。

这就是典型的三段论逻辑推理的模式，我想你会同意这里的逻辑推理是完全合乎情理的：如果所有的人类都是凡人，终有一死，而艾玛是人类之一，所以她也是凡人，也终有一死。

我们继续看下一个例子：

所有的教授颜值都高；

迈克尔是一名教授；

所以迈克尔是位帅哥。

很显然，教授们倒是很乐意相信这个结论。然而，从逻辑的角度来看，这个推论也是没有什么错误的：事实上，从推论本身来看，它完全正确。如果真的所有教授颜值都很高的话，迈克尔是一名教授，那么得出他是帅哥的结论就完全合乎情理。逻辑并不关心你一开始的陈述是否真实(即前提是否真实)，只关心你使用的模式和得出的结论是否合理。当然，那是在承认这个前提的基础上。

接下来我们看一个不完备的推理：

所有学生都努力学习；

索菲是一名学生；

所以索菲很有钱。

这个推理就是错误的，因为仅仅依靠前两个前提就得出索菲有钱的结论是不合理的。索菲有可能很富有，但这不是重点，重点在于你不能从给定的前提得出这个结论。

所以，逻辑是有关推理模式的，上述的例子所示的三段论也许是最简单有用的逻辑推理案例。逻辑告诉我们怎样正确地从前提中得出结论，这个过程被称为演绎。

三段论是古希腊哲学家亚里士多德(Aristotle)提出的，1000多年以来，三段论为逻辑分析提供了主要框架。然而，它能展示的逻辑推理形式十分有限，不适合许多复杂形式的论证。从很早开始，数学家就对理解推理的通用原理有着浓厚兴趣，因为数学的根本性问题都是有关推理的：数学家的工作就是从现有的知识中获取新的知识，换句话说，就是进行推理和演绎。到了19世纪，数学家们普遍对他们的工作原理感到困惑。他们想知道，到底什么是真实的？我们怎样证明数学论证是合理的推论呢？我们怎么确定1 ＋ 1 ＝ 2是正确的？

大约始于19世纪中叶，数学家们开始认真研究这些问题。德国的戈特洛布·弗雷格(Gottlob Frege)发展了普通的逻辑演算，为世人第一次展现了类似现代数理逻辑框架的东西。伦敦的奥古斯都·德·摩根(Augustus de Morgan)和来自爱尔兰科克城市的乔治·布尔(George Boole)展示了如何将应用于代数问题的相同计算方式应用于逻辑推理(1854年，布尔发表了相关论文，并起了个傲慢的题名：思想法则)。

到了20世纪初，现代逻辑的基本框架已经大致建立起来，当时确立的逻辑运算系统，直至如今仍然能够支撑数学家几乎所有的逻辑推理工作。这个系统被称为一阶逻辑，一阶逻辑是数学和推理的通用语言。这个框架涵盖了亚里士多德、弗雷格、德·摩根、布尔和其他人的所有类型的推理。在人工智能领域，一阶逻辑似乎以同样的方式提供了统一框架，为当时世界各地各式各样、不成体系的知识表述方案提供了统一框架。

基于逻辑的人工智能范式由此诞生，其最早和最具影响力的倡导者之一，就是约翰·麦卡锡——正是那位在1956年主持了达特茅斯暑期学校，并为人工智能命名的麦卡锡。他描述了自己对基于逻辑的人工智能系统的构想[36]：

我的观点是，智能体能够使用逻辑语句来表达其对世界、目标和当前状况的了解，并通过(推断)某个行为或者行动过程是否符合它的目标，来决定它自己的行为。

这里的“智能体”指的就是人工智能，麦卡锡提出的“逻辑语句”正是我们前文所述的类似“凡人皆有一死”或者“所有教授颜值都高”等陈述。一阶逻辑提供了丰富的、数学上能够精准表达的语言，可以用来表述逻辑语句。

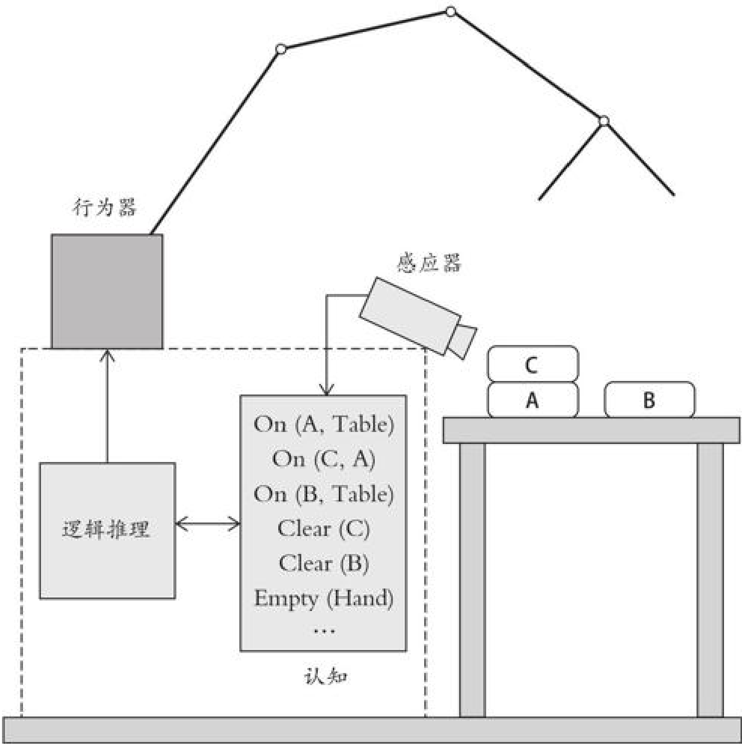


图8　麦卡锡对人工智能的构想

麦卡锡设想的基于逻辑的人工智能，一个基于逻辑的人工智能系统能通过逻辑语句明确地表示其对环境的信念，智能决策则被分解为逻辑推理的过程。

图8展示了麦卡锡对逻辑人工智能的构想(相当程序化)。我们看到一个机器人在类似积木世界的环境中工作，它配有机械臂和用于获取环境信息的感应系统。在机器人内部，有关于环境的逻辑描述，包括On(A，Table)， On(C， A)等语句。这些语句实际上是一阶逻辑的表达式。对于图8所示场景，它们的含义非常明显：

·On(x， y)表示对象x位于对象y的上方

·Clear(x)表示对象x上没有对象

·Empty(x)表示x是空的

在基于逻辑的人工智能系统中，关键任务之一是定义表达式词汇表。比如On(x， y)、Clear(x)等等，用来表示所处的场景。

机器人内部的逻辑推理系统用于判断它从所处环境中获取的所有信息，这种认知通常被称为信念。机器人的信念系统包含了“On(A，Table)”就意味着“机器人相信对象A在桌面上”。不过，我们需要谨慎地对待这些术语。虽然使用我们熟悉的词汇来做术语有助于我们理解机器人的“信念”〔对人类而言，“On(A，Table)”是一种非常容易理解的表述〕，但对机器人而言，只有它的知识库里存储了这个表达式的意义，才能和机器人本身的行为产生关联。只有机器人的知识库中有了这一条认知以后，才真正算作一项“信念”。另外，选择“On”作为术语也没什么神奇之处，其实设计者完全可以使用诸如“Qwerty(x， y)”之类的语言来表述物体x位于物体y之上，对机器人而言，其执行结果并没有什么不同。人工智能开发中一个常见的误区，就在于开发者为系统所选择的术语对人类而言太容易理解了，反而增加了许多不必要的误解，这让我们以为机器人也能轻而易举读懂语言本身，而事实并非如此。

机器人的感应系统负责将感应器所提供的原始信息转化成机器人能使用的内部逻辑形式，如图8所示。正如前面我们提到过的，这不是一件容易的事情，我们稍后会详细讨论。

最后，通过机器人的逻辑推理系统(推理机)来决定机器人的实际行为。逻辑人工智能的核心思想就在于机器人的行为是基于逻辑推理的结果：它必须推断出自己接下来应该做什么。

总的来说，我们可以设想机器人的行为模式：在一个周期内不断地

通过传感器观察周围环境，更新对环境的认知，推断出下一步该做什么，执行这个动作，然后重新开始整个过程。

基于逻辑的人工智能为什么会有相当的影响力？我想我们有必要理解这一点。或许最重要的原因是它让一切变得单纯：构建智能系统的整体问题被简化成对机器人应该做什么的逻辑描述。而这样一个系统是透明的：要理解机器人为什么这么做，只需要看到它的信念和推理过程就行了。

当然，我认为还有一些不那么明显的原因。首先，想象一下，我们做决定的依据来源于推理，这个想法本身就很吸引人。当我们思考的时候，似乎也是这么做的：我们考虑各种行动方案的优劣，可能会和自己来一场心灵对话，厘清每种方案的利弊。而基于逻辑的人工智能似乎遵循了这一过程。

逻辑编程

从20世纪70年代末到80年代中期，基于逻辑的人工智能范式开始逐步拥有影响力，到了80年代初，它已经成为人工智能的主流。研究人员开始推测，它不仅在人工智能领域有用，甚至可以推广至整个计算机领域。逻辑编程开始广为人知，从根本上改变了人们的编程模式。长期以来，编程是一项烦琐、耗时且容易出错的工作，因为它迫使人们思考计算机程序操作的每一个步骤，并且不能出一丁点儿错，人应付起来相当艰难。而逻辑编程把程序员从这样的诅咒里解脱出来，这是它的最大卖点。在逻辑编程中，你可以利用逻辑的力量来表达你对问题的了解——剩下的事情，交给逻辑推理即可。你没必要详细写出每一个步骤，一个逻辑程序会推断出它实际上需要做什么。

逻辑编程使用的是一种非常有名的语言：**PROLOG**[37]，它可能算是逻辑人工智能时代最卓越的遗产了——这种语言至今仍在广泛使用和传授。PROLOG主要是由美籍英国研究员鲍勃·科瓦尔斯基(Bob Kowalski)，以及两名来自法国马赛的研究员，阿兰·科尔默劳尔(Alain Colmerauer)和菲利普·罗塞尔(Philippe Roussel)发明的。在20世纪70年代早期，科瓦尔斯基就意识到可以使用一阶逻辑规则来构建编程语言的基础，虽然他有了这个新想法，但是并没有落实细节——这项工作是在1972年科瓦尔斯基拜访之后，由科尔默劳尔和罗塞尔完成的。

PROLOG是一门非常直观的语言，我们来看一个例子。下面是如何用PROLOG表示前文举的“凡人皆有一死”的例子：

```
human (emma).
mortal (X): - human (X).
```

第一行程序是PROLOG中定义的“事实”，即艾玛是凡人。第二行是PROLOG中的“规则”，表示若是X是凡人，X终有一死。

在实际中运用PROLOG，我们需要给它一个目标。在本例中，我们需要确认艾玛是否会死亡，可以这样表示：

```
mortal (emma).
```

这句语句的目的是向PROLOG询问“艾玛是否终有一死”，或者用更具体的方式表述，即“你能根据已知的事实和规则推断艾玛是否终有一死吗？”有了这个目标，PROLOG就能使用逻辑推理来证明，该语句为真[38]。(有关PROLOG的深入了解，请参阅附录B。)

上述例子只展示了PROLOG强大功能的冰山一角，对于某些问题，你可以用PROLOG编写出非常简洁优雅的程序。1974年大卫·沃伦(David Warren)编写的战机计划系统可以解决积木世界内的规划问题，只需要100行PROLOG代码[39]。如果要用Python之类的语言编写同样的系统，大概需要数千行代码以及数月的工作。此外，用PROLOG编写的程序不仅仅是程序，还是逻辑公式。因此，PROLOG程序看上去实现了麦卡锡对基于逻辑的人工智能的设想。

在20世纪70年代末到80年代初，PROLOG开始崭露头角，直到它挑战麦卡锡倍加推崇的LISP语言——人工智能黑客首选的编程语言。80年代初，PROLOG成为日本政府大规模投资的对象，被称为第五代计算机项目。当时，与美国和欧洲的经济相比，日本经济呈现蓬勃发展之势，这也让美国和欧洲政府十分担心。但同时，日本政府也有所顾虑，虽然日本工业擅长技术再开发和商业化，但在根本的创新性方面，尤其在计算机领域，它并不十分出色。第五代计算机项目旨在改变这一现状，试图确立日本全球计算机创新中心的地位，而逻辑编程是定向投资的关键技术之一。在后续的十年里，大约4亿美金被投入到第五代计算机项目中。不过，尽管它在建立日本的计算机科学基础方面发挥了重要作用，却并没有让日本的计算机产业超越美国。PROLOG也没有像逻辑编程界

人士所期望的那样，成为一种通用计算机语言。

虽然PROLOG没有征服程序员的世界，但不能将它视为失败。毕竟，在世界各地，程序员每天都在愉快地使用PROLOG编写高效的程序，并感激这种语言的强大和优雅。但对我来说，PROLOG最大的贡献却是无形的：逻辑编程为计算机提供了一个全新的思路，就如何解决计算问题而言，它有着跟传统计算机完全不同的观点。一代又一代程序员因此受益匪浅。

Cyc：终极专家系统

Cyc工程可谓知识型人工智能时代最著名的实验了，Cyc工程是天才的人工智能研究员道格·莱纳特(Doug Lenat)的智慧结晶，莱纳特拥有卓越的科学能力、坚定不移的信心以及优秀的说服能力，让旁人能够接受他的愿景。在20世纪70年代，他凭借一系列令人印象深刻的人工智能系统崭露头角，1977年，他荣获计算机与思想奖，这是能够授予年轻人工智能科学家的最具声望的奖项。

在80年代初期，莱纳特确信“知识就是力量”理论的应用应该远远超过大家所关注的、狭义的专家系统。他确信，这个理论可以成为打开通用人工智能——人工智能的宏伟梦想的钥匙。他在1990年写下这样的话〔合著者为拉曼内森·古哈(Ramanathan Guha)〕[40]：

> 我们不相信任何通往智能的捷径，世界上不存在尚未被发现的麦克斯韦思想方程式之类的通解……任何强大的思维体都无可避免地需要大量知识。我们所说的知识，不是指枯燥的、专业领域内的知识；更确切地说，是现实世界中的各种常识……它们太普通了，所以参考书里面找不到。例如：生命不可永恒，任何实体不可能同时出现在两个不同的地方，动物不喜欢承受痛苦，等等。恐怕人工智能所面临的最艰难的事实——过去34年来人工智能领域一直致力于摆脱的事实——其实是，要获取如此庞大的知识库，没有一种优雅、轻松的方式。相反，大部分工作(至少在初期)必须经过人类判断以后手工输入。

因此，在80年代中期，莱纳特和他的同事给自己立下了艰难的挑战任务，创建“包罗万象的知识库”，Cyc工程应运而生。

Cyc工程的雄心壮志确实令人难以置信，要实现莱纳特的构想，Cyc的知识库需要对“共识现实”，即我们所理解的整个世界进行完整描述。现在让我们暂停阅读，思考一下这个挑战有多难，也就是说，必须有人明确告诉Cyc很多对人类而言理所当然的常识，比如：

> 在地球上，一个悬在空中的物体会落地，它落到地面

时会停下来；而在太空里面，悬空物体则不会落地。

燃料耗尽的飞机会坠机。

坠机事故会让人丧生。

食用你不认识的蘑菇是危险行为。

红色水龙头往往通的是热水，蓝色的往往是冷水。

……

一个受过良好教育的人在生活中所认知的日常知识，在Cyc系统里面必须由人用特殊的语言明确地写下来，并且输入系统中。

莱纳特预估这个输入的工程需要至少200人经过多年的努力来实现，大部分的时间和人力都会消耗在手动知识录入上，即告诉Cyc我们的世界是怎样的，以及我们是怎么理解它的。莱纳特乐观地认为，用不了多久，Cyc就可以实现自我学习。一个以莱纳特设想的方式自主教育、自我学习的系统，就意味着通用人工智能的设想成为现实。这就是说，莱纳特假设通用人工智能的问题主要是知识储备问题，它可以通过一个合适的基于知识的系统来解决。Cyc假设是一场豪赌，高风险、高收益，如果被证明成功，它就能改变整个世界。

科研界，包括科研投资的悖论之一就是，有时候，宏伟的目标会掩盖其荒谬的本质。事实上，正是某些研究机构不遗余力地支持和鼓励那些雄心勃勃的、步子迈得太超前的科研者，而摒弃了一些安全但是进度缓慢的研究。也许Cyc就是这样一个案例。不管怎么说，1984年，Cyc工程从得克萨斯州奥斯汀市的微电子和计算机联合会那里获取了资金，项目启动了。

当然，它失败了。

首要的问题在于，以前从来没人试图将人类所有的基础常识组织汇编，你到底该从哪里开始？第一步，你得定义将使用到的所有词汇，还有各种术语，它们之间是怎么关联的。定义基本术语和概念，并围绕它们组织整个知识体系，被称为本体工程。而Cyc的本体工程面临的挑战，比此前所有的其他项目都庞大得多。很快，莱纳特和他的同事就发现，他们关于如何构建Cyc的知识体系，以及如何获取知识的想法都太过天真和粗糙。并且，随着项目的推进，时不时还得推翻重来。

尽管经历各种挫折，在项目实施10年后莱纳特仍然保持乐观。1994

年4月，他的乐观预期引起了斯坦福大学计算机科学教授沃恩·普拉特[41] (Vaughan Pratt)的注意。普拉特是一名颇有威望的计算理论专家，习惯用精准而严谨的数学语言来思考计算机的本质。普拉特对一次Cyc的展示会颇感兴趣，他要求现场观看一次程序演示。莱纳特同意了。

在访问之前，普拉特给莱纳特写了封信，根据公开发表的有关Cyc项目的科学文章，普拉特列举了一个清单，包含他期望Cyc能够实现的功能。普拉特写道："我认为，如果我的期望和Cyc的能力相匹配，那么这个演示程序将发挥最佳效果。如果我的期望太高了，那肯定会令我失望。但如果我低估了它……我们可能花费太多时间在不公正地质疑这个项目上了。"但他没有收到回复。

1994年4月15日，普拉特见到了Cyc项目的演示。演示从一些很简单的例子开始，展示了Cyc系统看似不够亮眼但真正富有技术含量的成就——识别出数据库中需要Cyc运用推理能力才能识别的细节差别。真是一个良好的开端。然后莱纳特开始演示一些扩展能力，让Cyc识别出用自然语言编写的需求，按照需求检索图片。"放松的某人"的查询结果是三个穿着沙滩服的男人拿着冲浪板的照片——Cyc正确地将冲浪板、冲浪和放松联系起来。普拉特简略书写了一下Cyc推理到这个环节的推理链，它需要运用近20条规则，其中某些规则在我们看来可能挺奇怪的——"所有哺乳动物都是脊椎动物"是完成这个推理链所用到的规则之一(普拉特看到的Cyc版本有50多万条规则)。

到目前为止进展得不错，不过，普拉特想了解Cyc最核心的特征之一：它对世界的实际认知。然后，普拉特询问Cyc是否知道面包是食物。主机将问题翻译成Cyc使用的规则语言。"是的，"Cyc回答说。"那么，Cyc认为面包是一种饮料吗？"普拉特问道。这次Cyc卡住了。他们试图明确地告诉它面包不是饮料。普拉特在报告里写道："在稍加调整之后，我们放弃了这个问题。"他们继续进行演示，Cyc似乎知道各种活动会导致死亡，但并不知道饥饿也是死亡的一个原因。事实上，Cyc似乎没有"饿死"这个概念。普拉特继续询问有关行星、天空和汽车的问题，但很快就发现，Cyc的知识储备十分粗略和难以捉摸。例如，它不知道天空是蓝色的，也不知道汽车通常有四个轮子。

在观看过莱纳特的演示以后，普拉特拿出了准备好的，包含100多个问题的清单，他希望这些问题能够探索Cyc到底对这个世界的常识了解多少，理解到何种程度。例如：

如果汤姆比迪克高3英寸[6]，迪克比哈里高2英寸，那么汤姆比哈里高多少？

两兄弟可以彼此都比对方更高吗？

陆地和海洋哪个更湿？

汽车能向后开吗？飞机能向后飞吗？

人没有空气能活多久？

这些都是关于这个世界最基础的常识性问题，但Cyc不能对此做出有意义的回答。普拉特回顾了自己的经历，对Cyc项目提出了谨慎的批评。他推测，或许Cyc已经朝着通用人工智能的目标跨出了一步，取得了一定进展，但问题在于，很难弄清楚这个进展有多大，因为Cyc的知识体系非常零碎，分布得相当不均匀。

我们能从Cyc项目中获得哪些有关通用人工智能道路的启示呢？忽略过于乐观的预期，其实在大规模知识工程中，Cyc算是拥有复杂技术的一次实践。它虽然没有给出一条指向通用人工智能的光明大道，但它教会了我们在开发和组织基于大型知识基础的知识系统中的各种方法。从严格精准的意义上来说，**Cyc**假说——通用人工智能的本质是知识体系问题，可以通过一个合适的基于知识的系统来解决——既没有被证实，也没有被证伪。Cyc不能成为通用人工智能成功案例并不能证明这个理论是错误的，只能证明这种方式没有用而已。可以想象或许存在更合理的基于知识的体系，能够带来更大的进步。

从某种意义上说，Cyc项目是领先于时代的。Cyc项目开始30年以后，谷歌发布了知识图谱，这是一个庞大的知识库，知识图谱里面汇集了大量关于实体世界的信息(地点、人物、电影、书籍、重大事件等)，这些信息被用来丰富谷歌搜索引擎的查找结果。当你使用谷歌搜索引擎时，你所查找的东西通常会涉及一些实体的名称或者其他术语之类。举个例子，如果你搜索“麦当娜”，简单的网络搜索结果会返回包含“麦当娜”这个关键词的网页。但是，如果搜索引擎知道麦当娜是一名流行歌手，以及她的全名是麦当娜·路易斯·西科尼(Madonna Louise Ciccone)，就会返回一些更有价值的东西。所以，你在谷歌上搜索“麦当娜”就会发现，你的搜索结果包含了这位歌手的一系列信息，这些信息会回答人们有关她的常见问题(比如出生年月、孩子状况、最受欢迎的专辑等)。Cyc和谷歌知识图谱最关键的区别在于，知识图谱中的知识并非手工编码的，而是自动从维基百科等网页中提取出来的。到2017年，据称维基百

科包含700亿条名词，涉及大约5亿个实体。当然，知识图谱并没有以通用人工智能为目标，目前也不清楚它涉及了多少基于知识系统的至关重要的世界观推理。不过，不管怎么说，我认为，知识图谱多多少少还是携带了Cyc的基因。

然而，不管我们回顾历史时如何尽力客观评价Cyc项目，可悲的事实总是无法改变。Cyc项目在人工智能发展史中之所以留下浓墨重彩的一笔，源于它是一个人工智能被过度炒作的极端案例，它完全辜负了人们寄予的不切实际的厚望。莱纳特的名字也被编入了计算机科学的民间笑料里。有个笑话说，“微莱纳特”应该作为一个科学计量单位，用来衡量某件事情的虚假程度。为什么要在莱纳特前面加上“微”？因为你找不到比一个“整莱纳特”更虚假的东西了。

破灭出现

就像许多理论上很美好的人工智能一样，基于知识的人工智能被证明实际应用非常有限。这种方式在一些我们认为荒谬琐碎的推理任务中，也遇到了困难。考虑一下常识推理任务，下面的场景给出一个典型例子[42]：

> 你得知翠迪是一只鸟，所以你得出一个结论：翠迪会飞。然而，翠迪是一只企鹅。所以你收回了之前的结论。

乍一看，这个问题在逻辑上应该很容易理解。看上去跟我们之前讨论的逻辑三段论没什么差别——我们的人工智能系统存在知识“如果x是一只鸟，那么x会飞”，那么，给出“x是一只鸟”的信息，它理所当然能推理出“x能飞”的结论。好吧，到目前为止没有任何问题。但我们现在得知，翠迪是一只企鹅，接下来会发生什么呢？我们需要撤销之前得出的“翠迪会飞”的结论。问题来了，逻辑无法处理这样的问题。在逻辑上，增加更多的信息永远不会消除你之前得到的任何结论。但是增加的信息(即“翠迪是一只企鹅”)确确实实使我们需要撤销之前的结论(即“翠迪会飞”)。

常识推理还有另外一个问题，在我们遇到矛盾的时候就会凸显。以下是文献中的标准示例[43]：

> 贵格会教徒是和平主义者；
>
> 共和党人不是和平主义者；

尼克松是一名共和党人，也是一名和平主义者。

试着从逻辑上解释上述文字，你就会陷入理解困境，因为最终我们会得出结论：尼克松既是和平主义者，又不是和平主义者——这是一个矛盾。在这样的矛盾面前，逻辑完全失效：它不能应对这样的场景，也无法告知我们任何有意义的结论。然而，矛盾是我们每天都会遇见的现实。我们被告知纳税是有益处的，我们也被告知纳税是一种损失；我们被告知葡萄酒有益，我们也被告知葡萄酒有害；等等。矛盾在生活中无处不在。问题在于，我们试图用逻辑去处理它不该处理的领域——在数学中，如果遇见矛盾，就意味着你犯了错。

这些问题看似微不足道，但在20世纪80年代末，它们是基于知识的人工智能研究最主要的挑战，吸引了该领域最聪明的一群人，但从来没得到过真正意义上的解决。

事实证明，构建和部署专家系统比最初想象的困难得多。最主要的难题是后来被称为知识获取的问题，简单地说，就是如何从人类专家那里提取知识并以规则形式编码。人类专家通常也很难表达他们所拥有的专业知识——他们擅长处理某种事情并不意味着他们可以讲清楚实际上是怎么做到的。另外，专家们也并非那么渴望分享自己的专业知识。毕竟，如果你的公司能够用程序来代替你工作，他们还要你干什么？何况人工智能会取代人类的担忧一直存在。

到了20世纪80年代末，专家系统的繁荣已经结束。基于知识的系统技术倒不能说是失败了，毕竟许多成功的专家系统也在这一时期内建立起来，此后也逐步有更多的专家系统出现。但是，看样子，又一次证明，人工智能的实际产出没有达到被鼓吹炒作的高度。

第四章 机器人与其合理性

行而不思者，因莽撞而灭亡；

思而不行者，因懈怠而灭亡。

——威斯坦·休·奥登(Wystan Hugh Auden)

哲学家托马斯·库恩(Thomas Kuhn)在其1962年出版的著作《科学革命的结构》一书中指出，随着科学认知的发展，有时候旧有科学体系会面临全盘崩溃的危机，新的科学体系诞生，取代传统的、既定的科学体系，这就意味着科学的范式将发生变化。到了20世纪80年代末期，专家系统兴盛的时光日渐式微，一场新的人工智能革命悄然临近。诚然，人工智能仍然因期望太高、过度炒作、实现困难等问题为人诟病，但这一次，被动摇的不仅仅是专家系统的根基——“知识就是力量”学说，更是自20世纪50年代以来支撑人工智能，尤其是符号人工智能发展的基础设定。说来有趣，在80年代末期，对人工智能领域批评得最激烈的人士，恰好是来自该领域本身。

澳大利亚机器人学家罗德尼·布鲁克斯(Rodney Brooks)，是彼时对人工智能范式批评得最激烈，也是最富影响力和权威的专家。出生于1945年的布鲁克斯曾在斯坦福大学、麻省理工学院和卡内基-梅隆大学这三大人工智能研究中心学习和工作，照理说不可能成为人工智能的批判者。布鲁克斯梦想着制造能够在现实世界中执行有用任务的机器人，在20世纪80年代初期，他就开始质疑当时流行的一种机器人理论——制造机器人的关键是将现实世界的知识编码成某种可以被机器人识别的形式，作为推理和决策的基础。80年代中期，他在麻省理工学院担任教员时，开始重新思考人工智能最基础层面的问题。

布鲁克斯革命

要理解布鲁克斯的观点，我们得重新回到积木世界。回想一下，积木世界是一种由各种模拟元素组成的桌面，桌面上堆叠着不同的对象——人工智能的任务则是以特定的方式重新排列对象。乍一看，积木世界作为人工智能的试验场是完全合情合理的：它听起来像是一个丰富的仓库环境，我敢肯定，多年来人工智能研究的诸多提案中也明确提到了这一点。但对布鲁克斯以及他的追随者而言，积木世界是没有意义的。

原因很简单，它的积木只是模拟元素，而现实世界中的问题比模拟元素组成的积木复杂得多。能够在积木世界里解决问题的系统，无论它看起来有多么智能，回归到现实的仓库环境中，都毫无价值。因为在物理世界中，人工智能面临的真正困难是处理感知问题，而在积木世界中，感知问题完全被忽略。(在20世纪70至80年代，积木世界最终被确认为人工智能研究谬误的发展方向，然而这并没有使人们停止对积木世界的研究，即使在今天，我们仍然可以找到各种有关积木世界的研究论文。我承认，我自己也写过几篇。)

布鲁克斯提出了三个关键原则，确定了他的理论体系：首先，他确信，人工智能要取得有意义的进步，只能通过与现实世界中的系统互动来实现——它们必须直接处于某个现实环境，感知并与之互动。其次，他认为，不管是以知识还是以逻辑为基础的人工智能，清晰而全面的知识储备及推理并不是它们智慧行为产生的必要条件，尤其是以逻辑为基础的人工智能。最后，他认为，智慧是一种涌现性质[7]，来源于实体与它所处环境发生的各种交互行为。布鲁克斯用了一个发表于1991年的讽刺寓言来阐明自己的观点[44]：

假设19世纪90年代爆发了研究人工飞行(Artificial Flights， AF)的热潮，科学界、工程界和风险投资界都疯狂投入其中。一群AF研究人员奇迹般地乘上了时间机器，穿越到20世纪80年代，在一架商用波音747飞机的客舱里度过了难忘的几个小时。回到19世纪90年代，他们感到无比兴奋和狂热，毕竟确定了空中飞行是能够实现的。于是，他们立即着手研究所看到的一切，在设计倾斜座椅、双层玻璃方面取得了巨大的进步。AF研究员们坚信，只要研究出那些神奇的“塑料”，就能在人工飞行领域登顶称王。

这个寓言的重点在于，当我们思考人类智慧有关问题的时候，往往更关注那些迷人的、具象的方面，比如推理、解决问题或者下棋等智力活动。这些是学术界重视并且希望人工智能可以擅长的领域——相关研究也因为这样的偏向性而走了弯路。布鲁克斯和他的团队认为，推理和解决问题的能力或许会在智能行为中起到作用，但它们并不是研究如何构建人工智能的正确起点[45]。

回到布鲁克斯的寓言中，他还对“分而治之”的设定提出了异议，这一设定是自人工智能诞生之初就默认的基础：人们认为应该将人工智能行为分解为各个组成部分(推理、学习、感知)来研究，而忽略了这些组件如何协同工作：

(人工飞行研究人员)一致认为，这个项目太大了，不可能由单一团体完成，而且他们必须成为不同领域的专家。毕竟，在穿越的那几个小时里面，他们询问过同机乘客，波音公

司可是要雇用超过6000人来制造这么一架飞机呢。

……每个人都忙得不可开交，但是小组之间并没有太多交流。

最后，他指出那群研究人员天真地忽视了“重量”问题，他在寓言中写道：

……制造座椅的人用了最好的实心钢作为骨架，有人就犯嘀咕了，应该用空心钢管做啊，这样明显会减轻不少重量。可是众所周知，如果这么一架又大又重的飞机能够在天上飞行，这点重量只是九牛一毛。

这里的“重量”，布鲁克斯是在暗喻计算量。特别是，他强烈反对将所有的决策过程都简化成逻辑推理这种需要大量消耗计算机处理时间和内存的想法。

布鲁克斯为这则寓言选择的题目是“无表征智能”。作为一名研究人工智能的大学生，在20世纪80年代中期专家系统蓬勃发展的阶段，我所接受的教育都认为逻辑推理才是人工智能的核心。布鲁克斯的文章似乎从根本上否定了我以为自己所知道的有关这个领域的一切，感觉就像是异端。1991年，一名年轻的同事从澳大利亚大型人工智能研讨会上回来，兴奋地瞪大眼睛告诉我，斯坦福大学(麦卡锡故乡学院)和麻省理工学院(布鲁克斯所在学院)的博士生之间展开了一场激烈的辩论。一方坚持既定的传统——逻辑、知识表述和推理；另一方则倡导新的人工智能运动，他们不仅无礼地、公然地背弃了神圣的传统，还大肆地嘲笑它。

布鲁克斯或许是新研究方向知名度最高的倡导者，但他并非孤军奋战。也有许多研究人员得出类似结论，虽然在细节上有所偏差，但在不同的研究方法得到的最终的结果相当一致。

其中最重要的一点，是将知识和推理从人工智能的核心角色中抹去。在麦卡锡的人工智能系统设想中，逻辑模型是其核心，所有的人工智能行为都围绕这一核心进行(如上一章的图8所示)，而这个设想遭到了强烈抨击。当然也有一些比较中立的观点，认为推理和知识表述仍然可以在人工智能系统中发挥一定作用，虽然可能不是核心和主导作用。但更极端的声音认为应该完全摒弃它们。

这一点值得详细探讨。我们来回顾一下麦卡锡的逻辑派人工智能系统遵循的特定循环：感知其所在环境，推理其该做什么，然后采取行动。但是，以这种方式运行的系统，与环境是分离的。

现在，放下手里的这本书，花点时间四下观察。你可能身处候机室、咖啡厅、火车车厢、家里，或者躺在阳光下的小河边。当你环顾四周时，并没有从环境中脱离，你的感知和行为是与环境融为一体并保持协调的。当然，在某些情况下，你会脱离眼前的环境，陷入沉思。但这种情况并非常规状态，只能算偶然现象。

问题在于，基于知识的人工智能并没有反映出这一点。假设我给你介绍一个按照麦卡锡模型的逻辑人工智能设计的机器人，它通过一个持续的“感知—推理—行为”循环模式运行，利用处理器处理和解析感应器接收到的数据，将感知数据更新到信念库，然后推理应该进行什么操作，再执行它选择的行为，然后再次启动决策循环。现在，我自豪地告诉你，我设计的机器人，不管在执行什么任务，总是会选择最优的执行方式。让我们考虑一下这个机器人在一段时间里的运行模式：

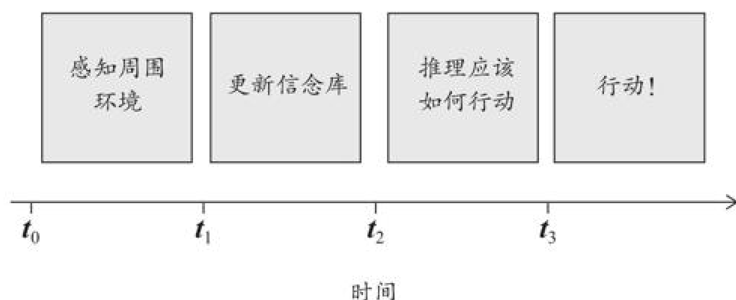


图9　机器人在一段时间里的运行模式

机器人应当选择最合适的行动时间，但是具体是什么时间呢？是感知周围环境的时候，还是最终决定执行什么操作的时候？

在t_0到t_1之间，机器人感知周围环境，然后用t_1到t_2的时间来处理传感器数据并更新信念库。在t_2到t_3之间，机器人对需要执行的动作进行推理，最后，t_3才是机器人开始行动的时间点。

现在我宣称自己设计的机器人总是采取最优的执行策略，但是想出最优策略对应的时间呢？是t_1还是t_3？机器人所掌握的环境信息是在t_1之前的，但在t_3以后它才可以基于这些信息开始正式行动。这种方法其实

是不切实际的，但令人惊讶的是，直到20世纪80年代末，几乎所有的人工智能都还在沿用这样的方式：人工智能研究者一直致力于制造能够在理论上做出最佳抉择的机器(前提是假设外部环境不会发生变化，同时又能弄清楚它到底应该做什么)，而不是实践中的最佳决策[46]。

因此，当时出现了另一个关键性的研究主题，即人工智能系统所处的环境和它所表现的行为之间，应该存在一种紧密的耦合关系。

基于行为的人工智能

如果布鲁克斯仅仅是一个批判人工智能的学者，那他的想法不会这么吸引人。毕竟，到了20世纪80年代中期，人工智能从业人员早就习惯无视批评的声音，不顾反对地继续自己的研究工作。布鲁克斯之所以从众多步休伯特·德莱弗斯(即第二章里出现的，撰写《炼金术与人工智能》的美国哲学家)后尘的批评家中脱颖而出，是因为他用令人信服的方式建立了另一种范式，并在随后的几十年里用一些令人印象深刻的系统证明了它。

人工智能的新范式被称为基于行为的人工智能，正如我们接下来要讨论的，它强调了特定的个体行为的重要性，这些行为有助于智能系统的整体运行。布鲁克斯采用的特殊方法被称为包容式体系结构，在当时出现的所有方法中，它似乎是影响力最持久的。现在我们来看看使用包容式体系结构制造的一种低调但实用的机器人：智能扫地机器人[8]。扫地机器人需要在屋子里四处游走，避开障碍物，并且在探测到垃圾或者灰尘的时候进行清洁——当机器人电量不足或者垃圾收纳盒装满时，我们希望它返回到固定位置并且关机。

包容式体系结构最基础的步骤是识别出机器人行为所需要的单个行为组件，然后用逐步添加组件的方式创造机器人。该结构的关键难点在于思考这些组件行为是如何相互关联的，并用何种方式组织它们，让机器人在正确的时间里表现出最恰当的行为。这通常需要对机器人进行广泛的实验，以确保组件行为以合理方式组合工作。

我们的扫地机器人需要六种基本的行为组件。

·避开障碍物：如果我发现前进方向有障碍物，我会改变行进方向，随机选择一个新的方向行进。

·关机：当我返回到充电座并且电量不足的时候，我会关机。

·清空垃圾收纳盒：如果我在充电座上，并且垃圾收纳盒有东西，就清空垃圾收纳盒。

·返回充电座：如果我的电量不足或者垃圾收纳盒已满，就返回充电座。

·执行清洁：如果我在当前位置检测到灰尘污物，则吸入垃圾收纳盒。

·随机移动：随机选择一个方向，然后朝该方向移动。

接下来要解决的问题是如何将这些组件行为组织起来。布鲁克斯建议使用他命名为包容式体系结构的方式组织(见图10)。包容式层次的结构决定了行为的优先级，层次结构中的组件行为越靠近底层，就拥有越高的优先级。在扫地机器人中，避开障碍物是优先级最高的行为，如果遇见障碍物，机器人会第一时间选择避开，始终优先执行这个操作。不难看出，这六种基本的组件行为组成了图10中的层次结构，将用以解决机器人的实际问题：机器人搜索所在区域的污物，如果有所发现，只要它不处于低电量或垃圾收纳盒已满的状态，则将污物吸入垃圾收纳盒；若是电量不足或者收纳盒已满，它会返回充电座。

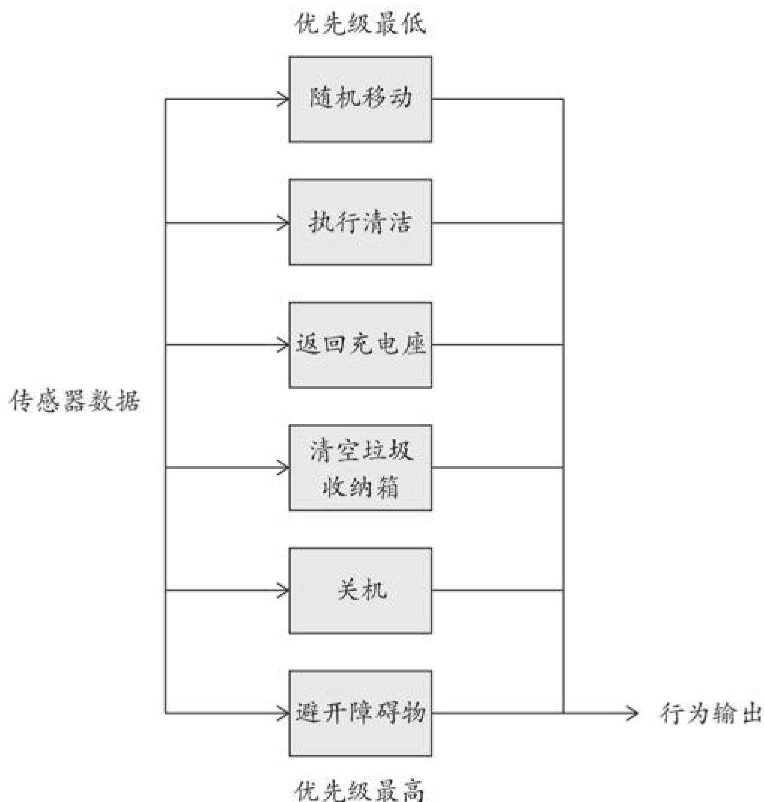


图10　扫地机器人包容式体系结构的简单示意图

虽然扫地机器人的行为看上去像是由规则控制的，但实际运作要简单得多。在机器人中实现这些行为，不需要像逻辑推理那样费劲。事实上，它们可以直接用简单电路的形式实现，这样一来，机器人将对传感数据的变化做出非常迅速的反应：它能快速响应环境变化。

随后的几十年，布鲁克斯使用包容式体系结构作为框架，开发了一系列令人瞩目的机器人。例如，他的“成吉思”机器人[47]，现陈列在美国国家航空航天博物馆中，外形就像六条腿的昆虫，采用包容式体系结构来组织57种基本的组件行为。如果用基于知识的人工智能技术来构造“成吉思”机器人，那会困难到令人惊讶——假设可以实现的话。在经历了几十年的边缘化之后，这些发展推动了机器人技术重新回到人工智能的主流。

基于智能体的人工智能

20世纪90年代初，我遇见了一位人工智能革命中的主角，他是我心目中的英雄人物。我很好奇他究竟如何看待那些他极力反对的人工智能技术——知识表述与推理、问题解决还有计划等。他真的相信这些技术在未来的人工智能中毫无作用吗？“当然不是，”他回答，“但是我不能赌上我的名声赞同它们的现状啊。”真是一个令人沮丧的回答，尽管事后看来，他也许只是想给一位天真的年轻研究生留下深刻印象。但不管这位教授所说的是否是肺腑之言，其他人肯定会认为，基于行为的人工智能也无法逃脱被寄予过高期望的泥潭。不久后，人们开始清晰地意识到，虽然基于行为的人工智能对人工智能领域的基础假设提出了重要更新，但它仍然有非常严重的局限性。

问题就在于它无法扩展规模。如果我们只是想做个扫地机器人，那么基于行为的人工智能就能满足需求。扫地机器人无须推理，也不需要用自然语言跟人交互，或者解决复杂的问题。因为不需要考虑这些问题，所以它能够使用包容式体系结构(或者其他类似结构——当年有不少与之相类似的方式)来获取有效的解决方案。不过，虽然基于行为的人工智能在某些问题上(主要是机器人技术)取得了极大成功，但它并没有为人工智能提供灵丹妙药。一旦基础行为数量太多，就很难设计出行为系统，因为理解各个行为之间可能存在的相互作用会变得非常困难。使用基于行为的方法构建系统是一门类似于黑魔法的艺术——真想知道这个

系统是否实用，唯一的方式是构建出来并使用它，这种实践方式不仅昂贵、耗时，而且不可预测。另外，使用基于行为的方法构建的解决方案，虽然能够针对某个非常具体的问题提出精准有效的方法，但从中所积累的经验，很难应用到新问题上。

布鲁克斯表示知识基础和推理等能力并不是构建智能行为必需的基础，这无疑是正确的，他的机器人也向我们展示了纯行为模式可以达到的高度。但我认为，若是因此机械地认为推理和知识表述在人工智能领域毫无用处，那就错了。有些情况下，推理是无法避开的(不管是逻辑推理还是其他形式的推理)，试图否认这一点，就像试图用逻辑推理来构造一个扫地机器人一样荒谬。

虽然新型人工智能的某些支持者态度强硬——严禁在人工智能中加入任何类似逻辑表达和推理的东西，但大多数人则采取了一种更温和的方式，这种温和路线似乎一直盛行至今。温和派吸取了基于行为的人工智能的主要优点，但他们认为正确的解决方式是将行为和推理的方法相结合。一个新的人工智能发展方向开始出现，它吸收了布鲁克斯学说的精华，同时也接受此前被证明在人工智能推理和知识表述方面取得成功的因素。

人工智能研究的焦点再一次转移，从专家系统、逻辑推理机等脱离实体的系统转向构建智能体(Agent)。智能体指的是一个完整的人工智能系统，它是一个独立的、自主的实体，嵌入某个环境之中，代表用户执行特定的任务。一个智能体应该能提供一套完整的、集成的能力，而不仅仅是类似逻辑推理那样孤立的、脱离实体的能力。重点在于，将开发目标专注于构建完整的智能体，而不仅仅是智能的组成部分。人们希望这样的人工智能可以避开那个激怒过布鲁克斯的谬论——人工智能可以通过分离智能行为的组成部分(如推理、学习等)，彼此孤立地开发，然后组合在一起，取得成功。

基于智能体的人工智能观点直接受到行为人工智能的影响，但又弱化了它的主旨。这一风潮在20世纪90年代初开始初现端倪，当时的人工智能界花费了不少工夫明确当我们提到“智能体”这个词的时候，确切的意指是什么。而一直到90年代中期，人们才达成共识，所谓的智能体应该具备以下三种特性：首先，它们必须反应灵敏，必须迅速适应自己的环境，并且能够在环境变化中适时地调整自己的行为；第二，它们必须积极主动，能够系统地完成用户赋予它们的任务；最后，智能体需要有

协作性，即在需要的时候能够和其他智能体合作。人工智能的黄金年代强调的是积极主动，即计划和解决问题；而基于行为的人工智能则强调反应灵敏的重要性，体现在适应所处环境并与之协调。基于智能体的人工智能要求两者兼而有之，此外，还向混合体中注入了一些新的东西：智能体必须和其他智能体合作。为此，它们需要社交技能——不仅是沟通技能，还有和其他智能体的协作、协调、谈判以推进任务完成的能力。

正是出于这种考虑——人工智能体需要社交化，使得基于智能体的人工智能范式从所有人工智能模式中脱颖而出。事后看来，人工智能界花费了这么多年才开始认真思考人工智能系统如何相互协作，以及当它们彼此协作时会出现什么问题，这似乎太奇怪了。虽然研究人员从图灵测试就开始强调社交能力，但它只是指代人工智能通过自然语言与人进行互动和日常对话的能力。在基于智能体的人工智能中，最受关注的并不是如何与人交流，而是如何跟其他的智能体一起协同工作。

图11为我们展示了一个典型的智能体设计。这个智能体被称为“旅行机”[48]，它的总体控制分为三个子系统：快速反应子系统的运行方式类似于布鲁克斯的包容式体系结构，它负责处理需要迅速响应并且无须推理的情况，例如避开障碍物；规划子系统负责规划如何实现智能体的目标；另外，建模子系统负责处理与其他智能体的交互。控制系统会听取三个子系统的建议，并决定遵循哪一个的。通常是非常直接的决定：如果快速反应子系统说“停止!”那么控制系统会迅速停止智能体的运动。在20世纪90年代初，人们开发了不少类似这样的智能体系统。

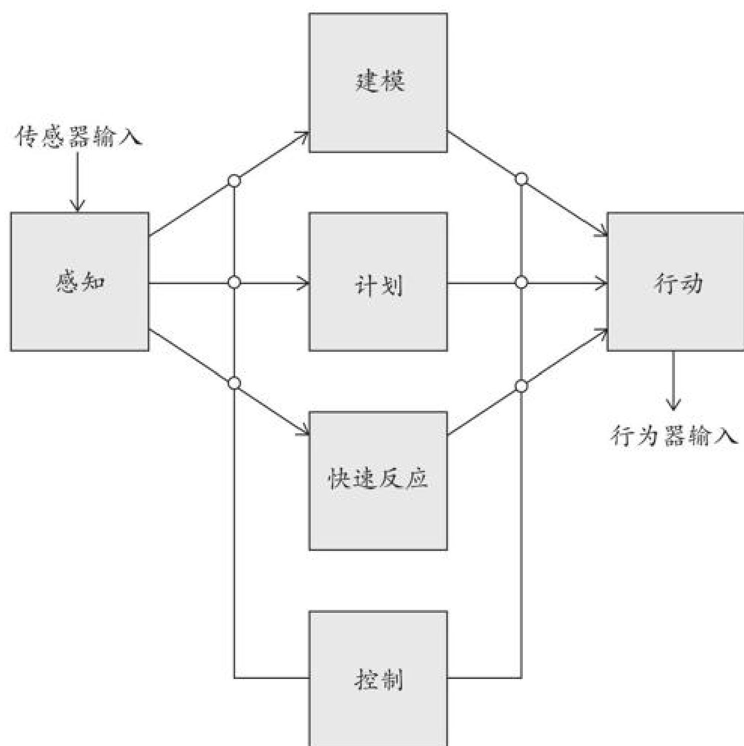


图11　一个典型的智能体结构：旅行机

HOMER：一个完整的智能体

HOMER是我最喜欢的例子，它是一种存在于二维“海洋世界”的模拟机器人潜水艇[49]，在海洋世界里完成一系列任务，通常负责收集包裹并将其从一个地点移动到另一个地点。显然，HOMER继承了SHRDLU的传统，并且有着同样的关键限制——它是一个模拟系统。但比起SHRDLU，它要复杂得多。

HOMER可以接受英语指令，不过用户指令的词汇表里只有800个单词。它可以计划如何完成指令，然后执行计划，并根据需要调整。为了了解HOMER的功能，我们来看看用户与它的对话。第一个问题涉及HOMER看到在它面前漂浮着的一根原木。(用户输入采用普通字体，HOMER的回复采用斜体字，我的注释放在括号中。)

你面前是什么？

一根原木。

它是属于你的吗？

不是。

它属于你。

哦。(HOMER用这个词表示被告知了以前不知道的事情)

……

转过身背对你的原木。

好的，我正在转身。

(用户把原木移除)

转回来。

好的，我正在转身。

我的原木不见了!

这段对话展示了一些简单但重要的常识性推理：HOMER记得那根原木，当它转身回来的时候，会期待看到它，当它发现原木不见了的时候，会表示惊讶。

下一个场景从HOMER在码头上等待任务开始，码头上有一个包裹，HOMER的任务是收集包裹并将其运送到驳船上。

下周六晚上9点把包裹放到驳船上。

好的。

(HOMER开始制订计划。)

你下周六会去码头吗？

是的。

你为什么要去码头？

我要去取包裹。

你打算什么时候去取？

晚上8点56分。

下周日包裹会在哪里？

在驳船上。

一周以后包裹会在码头上吗？

不会。

在这里我们看到HOMER展示了它对时间的常识性理解，以及它的行为是如何影响环境的——当它把包裹送到驳船上以后，包裹将不会出现在码头。它还表明了它理解自己的计划是需要时间来执行的——它需

要在晚上8点56分到码头取包裹，才能赶得上9点把它放到驳船上。

人工智能助手

虽然基于智能体的人工智能起源于机器人学，但许多研究人员很快意识到它在软件世界有着非凡的应用前景。没有人声称它是通用人工智能——就像专家系统的情况一样，人们只是想通过构建软件智能体来代替我们完成有用的任务。软件智能体的运行需要软件环境，如台式计算机和网络。最重要的想法是让人工智能驱动软件跟我们一起完成日常工作，就像助手一般能够处理电子邮件和上网。

为了弄明白软件智能体的概念是如何产生的，我们需要了解人类与计算机交互的方式，以及关于交互方式的思考是如何随着时间的推移而演变的。

最初的计算机用户，如艾伦·图灵，通常是帮助设计和制造计算机的科学家和工程师，他们制造的人机交互接口也就非常粗糙。图灵在20世纪40年代末使用的电脑“曼彻斯特宝贝”，其中每一个单独的计算机内存位置都必须通过开关来设置，以指示“0”或者“1”。计算机的一切都向用户公开，用户必须了解机器的工作原理才能编程。现在几乎没有人能掌握这种编程技术了，而即使在当时，如果你想换台机器编程，你从曼彻斯特宝贝身上学到的一切技术都毫无用处，因为编程的对象不同，程序也完全不同。

到了20世纪50年代末，这种情况开始发生改变，这一时期的主要创新是高级编程语言的开发，这些语言之所以被称为“高级”，是因为它们向程序员隐藏了一些机器语言的细节，因此程序员无须去了解某台计算机是怎样工作的，就能对其进行编程。这些高级语言独立于机器，从某种意义上来说，在一台计算机上用(比如)COBOL语言编写的程序可以在另一台计算机上运行——或许不一定能完美运行，或多或少出点错之类的，但编程技能开始可以中转了。这些创新极大地提高了计算机的可用性，标志着我们与计算机的交互方式开始从“以计算机为中心”转向“以人为中心”。

这种趋势在随后的60年一直在持续，从以机器为导向的人机交互模式稳步发展到越来越以人为中心的模式。到了20世纪80年代，随着苹果电脑公司于1984年推出麦金塔电脑(Mac电脑)，人机交互的方式又有了

巨大的飞跃。Mac电脑是第一台面向大众市场的电脑，它明确地表示不需要专业的计算机技能即可使用。它的主要卖点是用户界面，Mac拥有基于桌面的图形用户界面(GUI)，这就意味着Mac用户的界面上显示着代表文档和文件夹的图形图标(时至今日也是如此)，用户通过移动鼠标来操作这些桌面元素。虽然我们已经很熟悉用“桌面”这个词来描述自己的计算机界面，但你可能没想过，用这个词的原因是你的屏幕看起来像真实的桌面——你电脑上的文档应该像现实世界中的文档一样放在桌面上，文件夹也是现实中文件夹的类似物，可以帮助你整理文档。这样的类比还有：你像把废纸扔进垃圾桶一样把不要的文档或者文件夹拖进垃圾桶图标里……(现在这个比喻有点夸张了，我想很多年轻人只是模糊地意识到了这一点。)

1984年Mac电脑开创的图形用户界面，直至今日仍然是标准的[50]，而在硬件方面，从1984年到现在已经更新换代无数次了。我怀疑当年的Mac用户使用现代的Mac或Windows界面时不会存在任何困难。

因此，关于人机交互界面的要点是，用户需要使用他们熟悉的概念(如桌面、文档、文件夹、垃圾桶)与计算机交互，这能使得人机交互的过程更加自然，从而让用户访问计算机变得更加容易。

由于Mac的成功，在80年代末期，时任苹果公司CEO的约翰·斯库里(John Sculley)开始思考下一个类似Mac的人机交互创新可能是什么。他决定实现自己命名为知识导航器的想法，它预见了好几项新兴的计算机技术(尤其是万维网的出现，那时候距离万维网出现还有好几年时间)。为了说明知识导航器的想法，苹果公司委托制作了一段同名概念视频，于1987年发布[51]。这段视频是高度概念化的，意在展示一种愿景，而不是展示产品本身。

在视频中，我们看到一位大学教授，使用与现代的平板电脑高度相似的产品。这款平板电脑的界面看上去像是传统的桌面，但有一处关键的不同：教授与平板电脑的交互是通过软件智能体来实现的。与HOMER一样，这个智能体可以使用英文跟人进行交互，但与HOMER不同的是，该智能体在平板电脑的显示屏上是以动画人形的方式呈现的。智能体礼貌地提醒教授他的日程安排，查询一些有关教授正在计划的讲座材料，并负责管理来电。

这段视频别出心裁(其中包含了不少蹩脚的幽默)，但它却有着重要

的历史意义，原因有好几个。首先，它暗示了互联网将成为我们工作环境的常规部分。在视频制作的年代，互联网还没有普及到个人应用层面，甚至大多数公司都无法用到，它主要是学术机构和政府机关(尤其是军事机构)的专用品。另外，这段视频预示了平板电脑的普及。不过，对人工智能领域而言，视频最重要的一点是建立了通过智能体与计算机交互的思想。

基于智能体的交互界面代表了一种与从前截然不同的人机交互模式。当我们使用Microsoft Word或者IE浏览器之类的应用程序时，这些程序扮演的是被动的、只会响应的角色。这类应用程序不会控制，没有主动性。当你使用Microsoft Word的时候，它只会因为你选择了某个菜单或者点击了某个按钮而执行该操作，你和Microsoft Word之间的交互只有一个智能体，就是你自己。

基于智能体的交互界面则改变了这一点。计算机只能被动地等待用户告知它做什么，与此不同，一个智能体扮演的是更积极主动的角色，就像一个人类助手那样。智能体会以助手的身份与使用者合作，积极配合使用者做他想做的事情。

知识导航器视频中的智能体是以动画人物形式出现的，外形像视频会议中人物头像，用流利的英语发言。人工智能的能力远远超出了当时的技术，但这并不是核心思想的关键点，而是人工智能可以使软件成为使用者的合作者，不再是被动的仆人。这个中心思想随着视频深入人心，到了90年代中期，在万维网迅速扩张的刺激下，人们对软件智能体的兴趣迅速增长。

比利时裔麻省理工学院媒体实验室教授帕蒂·梅斯(Pattie Maes)发表了一篇广受读者欢迎的文章，文章的标题抓住了时代精神："智能体能够帮助人们减少工作量和信息过载[52]。"它描述了帕蒂实验室开发的许多原型智能体——电子邮件管理、会议日程安排、新闻过滤和音乐推荐等。比如，电子邮件助手会在收到电子邮件的时候观察用户的行为(立即阅读、将其归档、直接删除等)，并使用机器学习算法，尝试预测用户会怎么处理新到的电子邮件。当智能体对它的预测有足够满意时，就会主动行动，根据它的预测来为用户处理电子邮件。

随后的10年里，上百种类似的智能体问世，其中许多都基于互联网。在万维网早期，搜索工具还处于初级阶段，互联网连接比现在慢得

多。在网络上执行许多任务都非常耗时，人们希望智能体能够自动完成这些烦琐的工作[53]。对于20世纪90年代后期开始迅速发展的网络而言，软件智能体似乎是一种很有前景的技术，而一大批软件开发公司，也很快成了网络泡沫的一部分。

事实上，我也是这个故事的一部分。1996年夏天，几位同事邀请我加入伦敦一家雄心勃勃的初创公司，他们给了我三倍于学校的工资，我大概思考了半秒钟就同意了。我们的计划是使用智能体来加强网络搜索效率，我们希望能把万维网变成一个图书馆。但我们对商业一无所知，实话说，我们不懂怎么开发商业软件。很快我们就明白了一个残酷的真相：我们以惊人的速度烧掉投资者的钱，但真不知道怎么把它赚回来。那是一段痛苦的时光，我在加入公司9个月以后就离开了它，再过仅仅8个月后，公司倒闭了。

我那悲剧的创业经历也预示着几年后全球范围内所发生的事情，众所周知，互联网泡沫从1995年持续到2000年初。随着那些雄心勃勃、身价不菲的互联网初创新贵纷纷陷入资金窘境，未能实现盈利，2000年初，网络市场开始崩溃。

软件智能体只是互联网故事的一小部分，但它也是跟人工智能相关的最明显的部分了。其实人工智能研究者倡导的梦想并没有错，只是太超前而已。20年后，苹果公司iPhone智能手机的一款应用横空出世，Siri。它是由斯坦福国际研究院开发的——没错，就是30年前开发SHAKEY的同一机构。Siri是20世纪90年代软件智能体工作的直接产物，许多人工智能界人士立即将其与苹果公司的知识导航器视频联系起来。Siri的构想是一个基于软件的智能体，用户可以用自然语言与之交互，并且代替用户执行简单的任务。其他大众市场的应用商迅速跟进：亚马逊的Alexa、微软的Cortana和谷歌助手都实现了类似的功能。他们都将开发起源追溯到基于智能体的人工智能，当然，实际上它们不可能是20世纪90年代出现的，因为当时的硬件不足以支持它们运行。至少在2010年后，移动设备的计算能力才足以支持。

理性行事

基于智能体的范式提供了另一条有关人工智能发展道路的思考：构造能够有效代替我们行动的智能体。不过这又引发了一个有趣的问题，图灵测试确立了一个观点：人工智能的目标是产生与人类相似到无法分

辨的行为。但这跟智能体开发的思路不一样，其实我们只是想让智能体代替我们执行最优的选择，它的选择跟人类是否一样，那就无关紧要了。我们真正想要智能体做的是最正确的选择，至少尽可能做出最好的选择。因此，人工智能开发的目标从构建做出跟人类一样选择的智能体转向做出最优选择的智能体。

支持人工智能做最优决策的理论可以追溯到20世纪40年代，约翰·冯·诺依曼——就是我们在第一章认识的，为最早的计算机设计做出开创工作的诺依曼，他和同事奥斯卡·摩根斯坦(Oskar Morgenstern)发展了理性决策的数学理论。该理论表明如何将做出理性决策的问题转化为数学计算问题[54]。在基于智能体的人工智能中，智能体将用它们的理论为用户做出最佳决策。

智能体理论的出发点是用户的偏好。如果你的智能体要代替你做事情，那么它需要明白你的希望是什么。你当然想让智能体尽可能做出你喜欢的选择，那么，我们如何让智能体明白用户的偏好呢？假设，你的智能体要代替你在买苹果、橘子或者梨之间做出选择，它首先需要知道你对这三种不同结果的期望值。例如，假设你的偏好如示例一这样的：

橘子比梨好

梨比苹果好

在这种情况下，你的智能体在苹果和橘子之间做出选择，它选了橘子，你会很高兴；如果它选择苹果，你就会失望。这就是最简单的偏好示例，你的偏好关系描述了你如何对每一对备选结果进行排序。冯·诺依曼和摩根斯坦的理性决策需要偏好关系满足某些一致性的基本要求。例如，假设你的偏好是示例二这样的：

橘子比梨好

梨比苹果好

苹果比橘子好

这么看来你的喜好就有些奇怪了。因为从橘子比梨好、梨比苹果好能推断出你在橘子和苹果中更喜欢橘子，但这就和你的声明相矛盾。因此，你的偏好不满足一致性。这就让你的智能体没办法为你做出最优决策。

下一步就是将符合一致性的偏好进行赋值，使用被称为实体程序的

方式。实体程序的基本思想是为每一种备选项赋予一个数字值：数字越大，就代表偏好程度越高。例如，我们可以将橘子的偏好程度赋值为3，梨子为2，苹果为1，这样就可以描述前文第一个示例的情况了。因为3大于2，2大于1，这样的话，实体程序就能够正确地捕捉到第一例中的偏好关系。同样地，我们也可以用实体程序将橘子赋值为10，梨赋值为9，苹果赋值为0。在这种情况下，这个赋值的具体数值并不重要：重要的是赋值大小引起的结果排序。关键点在于，偏好设置必须满足一致性，才可以使用这种实体程序赋值的方式，用数值来表示偏好程度。看看前文所举的第二个示例，试试你能不能给苹果、橘子和梨赋值来表示这个偏好关系。

用赋值关系来表示偏好程度的唯一目的是使其可以用数学计算的方式做出最优选择。我们的智能体就可以选择偏好值最大的选择项，这就意味着它的选择可以达成我们最喜欢的结果。类似这样的问题被称为优化问题，在数学中得到了广泛的研究。

不幸的是，很多选择比这个复杂棘手得多，因为它们涉及不确定性。不确定性选择的设置会比较复杂，选择后的行为会有很多种可能性，我们所知道的仅仅是每一种结果出现的概率。

我们举例来说明这个问题，下面的场景，你的智能体必须在两个选项中做出选择[55]：

选项1：掷一枚硬币，如果是正面，你的智能体获得4英镑；如果是背面，你的智能体获得3英镑。

选项2：掷一枚硬币，如果是正面，你的智能体获得6英镑；如果是背面，你的智能体什么都不获得。

这种情况下，你的智能体应该选择1还是2？我认为选项1是更好的选择，但是，为什么呢？

为了理解原因，我们需要一个叫作预期效用的概念，此处的预期效用可以等价于在此选择下获得的平均收益。

所以，考虑到选项1。我们掷硬币的概率是对半的(不考虑正面和反面的细微重量差别)，所以我们预期正面和反面出现的次数平均下来应该是相等，即一半正面，一半反面。所以，你的智能体一半的时间会收到4英镑，一半的时间会收到3英镑。因此，你的智能体从选择1当中获得

的预期效用是$(0.5\times4)+(0.5\times3)=3.5$(英镑)。

当然，从实际上来说，你的智能体选择1的时候不可能得到3.5英镑的收益，只是如果选择的次数足够多，获得收益的平均值就是3.5英镑。

同样的道理，我们能计算出选择2的预期效用是$(0.5\times6)+(0.5\times0)=3$(英镑)，所以平均来说，选择2只能给你带来3英镑的预期效用。

冯·诺依曼和摩根斯坦的理论中，理性决策的基本原则就是会做出预期效用最大化的行为。在这种情况下，预期效用最大化的选择是选项1，因为它的预期效用为3.5英镑，大于选项2中3英镑的预期效用。

请注意，选项2中提供了诱人的获得6英镑的可能性，这比方案1中的任何结果收益都高，但是，将这个诱人的可能性与同样可能获得0收益的概率相权衡，就不难明白为什么选项1的预期效用比较高了。

预期效用最大化的想法经常被人们误解，有些人认为用数字计算人类的偏好和选择是一种令人厌恶的行为。这种厌恶通常来自一个错误的概念，即收益就等于金钱，或者预期效用最大化理论从某种意义上来说是自私的(因为假设一个使预期效用最大化的智能体行为是只考虑到自己的收益)。但收益这个东西不过是获取偏好数值的一个定义而已，冯·诺依曼和摩根斯坦的理论在对于个人偏好究竟是什么或者应该是什么这个问题上完全保持中立，这个理论同样也适用于天使和魔鬼的偏好。如果你是一心为别人牺牲的人，那也没关系，如果你的利他主义偏好被赋值表达，那么预期效用最大化理论同样适用于你，就如它也适用于世界上最自私的人那样。

到了20世纪90年代，构建能代表我们理性行事的人工智能的智能体范式——这里的理性来自冯·诺依曼和摩根斯坦的理性抉择模型——已经成为人工智能的新正统学说，时至今日仍然如此[56]。如果说有任何共同的主题将当代人工智能的各个分支结合起来，那就是这个。在当今几乎所有的人工智能系统中，都有一个数字收益模型，代表用户的偏好，并且系统将根据这个模型努力使预期效用最大化——代表用户理性决策。

应对不确定性

人工智能还有一个长期存在的问题——如何处理不确定性，在20世纪90年代这个问题变得尤为重要。任何一个现实的人工智能系统都必须处理好不确定性问题，有时甚至会处理许多。举个例子，无人驾驶汽车从传感器获取数据流，但传感器并非完美的，例如测距仪说“前方没有障碍物”，这个结论不能保证百分百准确。测距仪掌握的信息当然有重要价值，但是我们不能百分百信任它。那么，考虑到会出错的可能性，我们又该怎么利用它呢？

在人工智能的历史上，出现过许许多多应对不确定性方案的特别发明或者再发明，但到了90年代，一种特定的方法占据了主导地位，这种方法被称为贝叶斯推断。贝叶斯推断是由18世纪英国神学家、数学家托马斯·贝叶斯(Thomas Bayes)发明的，使用了同样由贝叶斯提出的贝叶斯定理。贝叶斯定理关注的是在新的信息面前，我们应该怎么理性地去调整既有认知。在无人驾驶汽车的案例中，认知与我们前面是否有障碍物相关，而新的信息就是传感器传来的数据。

另外，贝叶斯定理很有趣，因为它强调了人们在处理涉及不确定性的认知决策时有多么糟糕。为了更好地理解这一点，请考虑如下场景：

> 一种新的致命的流感病毒开始流行，感染概率为千分之一。人们研究出了一种新的流感检验方法，准确率为99%。你突发奇想去参加检测，结果呈阳性。

你该如何面对这个结论？

检测结果呈阳性，我想大多数人都会担心，毕竟，这个测试的准确率可是99%。所以，现在我问你，如果检测结果呈阳性，罹患该流感的概率是多少，我猜大多数人会说0.99(与测试的准确率一致)。

事实上，这个答案错得离谱。你罹患该流感的概率只有十分之一。这似乎违反了直觉，到底是怎么回事呢？

假设我们随机选择1000人做流感测试。我们知道实际感染概率，这1000人中大约有1个人会感染流感。所以我们假设这1000人中确实有1个人感染了流感，而999人没有。

首先我们考虑一下那个罹患流感的可怜人，由于测试准确度是99%，它将有99%的概率检测出患了流感的人。所以，我们可以预期这个结果是正确的(0.99的概率)。

现在来考虑那999个幸运儿，因为测试准确率为99%，所以大概每测试100个人就会有一个误诊。我们测试这999个幸运儿都是没有罹患流感的，预估大概有9～10份测试结果会出现错误。换言之，就是大概有9～10个人，明明没有感染，但测试结果呈阳性。

因此，如果我们对1000人进行流感测试，可以预计其中10到11人的检测结果是呈阳性的。但我们知道，只有一名检测结果呈阳性的人是真患病者。

简言之，由于流感的感染率很低，所以假阳性的被检测者远远多于真阳性(这就是医生为什么很不喜欢依靠单一检测做出诊断的原因之一)。

我们来看看贝叶斯推断在机器人学科中的应用。假设有一个机器人正在探索未知区域，比如一个遥远的星球，或者一片被地震摧毁的废墟之类。我们希望机器人能够探索未知区域并绘制地图。机器人可以通过传感器感知周围环境，但是有一个问题：传感器会错误报告。因此，当机器人进行观察的时候，传感器说“这个位置有障碍物”，可能是正确的，也可能是错误的，我们无法确定。如果机器人默认传感器数据总是正确的，那么它绘制的地图就会出现很大偏差，并且它还有可能根据错误的信息做出移动的决定并撞上障碍物。

这里可以像刚才的流感测试案例一样使用贝叶斯推断，我们利用传感器数据的正确率来更新机器人的信念，通过多次观察来确认障碍物位置，然后逐步完善地图绘制[57]。随着时间的推移，机器人对障碍物所在的确切位置的判断就会越来越精准。

贝叶斯推断的重要性在于为我们提供了处理不完美数据的正确方法：我们既不丢弃数据，也不全盘相信它是正确的。我们利用它来更新机器人的信念库，通过概率来确定信念库的正确性。

尽管贝叶斯推断的功能强大，但由于人工智能系统经常需要处理大量复杂且相关联的数据，所以要使得贝叶斯推断在人工智能领域得以应用，还需要做大量的工作。为了捕捉数据间的相互关联，人工智能研究人员开发了贝叶斯网络，简称贝氏网络，即用图像化的方式来表达数据之间存在的相互关联。贝叶斯网络主要来自朱迪亚·珀尔(Judea Pearl)的研究工作，她是一位非常有影响力的人工智能领域研究专家，在理解和

阐明人工智能中概率的作用方面作出了任何人都无法超越的贡献[58]。图12为我们展示了简单的贝叶斯网络示意，这个贝叶斯网络捕捉到了三个假设之间的关系：你得了普通感冒；你流鼻涕；你头痛。从一个假设到另一个假设的箭头表示它们之间的影响关系。粗略地说，从假设x到假设y的箭头表示假设x的真实性会对假设y的真实性产生影响。例如，你得了普通感冒会影响到你是否流鼻涕，如果你流鼻涕，你有可能得了感冒。这些不同概率之间的关系是通过贝叶斯推断来得到的。如果你想深入了解细节，请参见附录C。

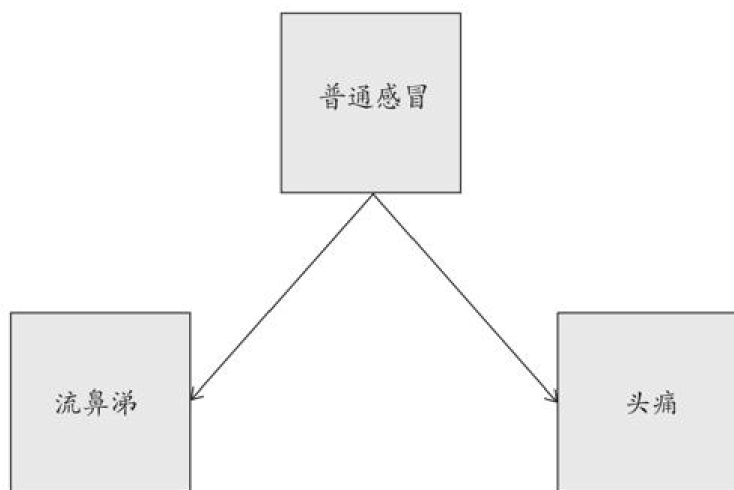


图12　简单的贝叶斯网络示意图

当Siri遇见Siri

虽然2000年的时候互联网热潮退却了，但研究者对智能体的兴趣仍然存在，人们开始关注智能体故事有可能出现的新转折点。研究人员想，如果智能体之间可以相互交流，那又会怎样？这个想法倒不是全新的：在以知识为基础的人工智能时代，研究人员就考虑过专家系统之间如何相互分享它们的专业知识，并开发出人工智能的语言让它们能够相互分享知识和查询对方所擅长的领域。但是在多智能体系统下，会出现一个关键性的差别：我希望我的智能体是尽力为我服务的，而你希望你的智能体是尽力为你服务的，既然你我之间存在兴趣爱好或偏好不一致

的情况，那么我们的智能体之间肯定也存在不一致。在这样的情况下，智能体就需要具备类似社交的能力。人们在日常生活中都会应用到这样的能力，人工智能所面临的新挑战就是构建具备社交能力的智能体[59]。

事后看来，之前开发者没怎么考虑过人工智能的社交属性似乎挺奇怪的。但是，在多智能体系统出现之前，人们的注意力都集中在开发单个的智能体上，而没有考虑到它会如何跟其他人工智能体交互。假设存在多个人工智能体，而不仅仅是一个，那就从根本上改写了人工智能发展的故事。智能体必须解决的问题是知道自己要做什么——该为自己所代表的用户做什么。如果一个智能体能够为我做出正确的选择，我会非常高兴。但如果周围存在其他智能体，那么我的智能体所选择的行为是好还是坏，至少在一定程度上取决于其他智能体的选择。因此，我的智能体在做选择的时候必须考虑到其他人的智能体会如何选择，同样，其他人的智能体也必须考虑到我的智能体的行为。

智能体在做决策的时候需要考虑对方的偏好和可能的行动，这样的推断实际上属于博弈论的研究范畴。博弈论以前主要是研究战略决策的经济学分支[60]，但正如其名所示[9]，博弈论起源于对象棋和扑克牌之类游戏的研究。其实，在多个智能体同时存在的情况下，它也是非常有实用价值的。

或许博弈论中最著名的中心思想，也是形成多智能体系统中决策基础的思想，就是纳什均衡。纳什均衡的概念是由小约翰·福布斯·纳什提出来的，我们在1956年被邀请参加约翰·麦卡锡达特茅斯人工智能暑期学校的学员名单中能找到他的名字。正是因为纳什均衡的提出，他与约翰·哈萨尼(John Harsanyi)和理查德·塞尔腾(Richard Selten)一同被授予1994年诺贝尔经济学奖。

纳什均衡的基本思想是很容易理解的。假设我们有两个智能体，每个都需要做出选择。智能体1选择x，智能体2选择y，那么我们如何判断它们是否做出了正确的选择呢？如果两个智能体都不后悔自己的选择，那它们的决策就是好的(从技术上讲，它们的决策形成了纳什均衡)。即：

鉴于智能体2做了选择y，智能体1满足于它所做的选择x；并且，鉴于智能体1做了选择x，智能体2满足于它所做的选择y。

纳什均衡之所以被称为均衡，是因为它捕捉到了决策过程中的稳定性：两个智能体都没有任何动机去做别的选择。

多智能体系统研究迅速采用了如纳什均衡的博弈论的思想作为系统决策研究的基础，但是一个熟悉的难题出现了。当纳什在20世纪50年代提出这个获得诺贝尔奖的想法时，他并不关心如何计算纳什均衡。不过，如果我们想让智能体在做决定时能够使用这个想法，这就一定是一个必须解决的重要问题。而且，也许可以预见，纳什均衡很难计算。寻找有效的方法来计算纳什均衡仍然是当今人工智能的一个主要课题。

人工智能开始成熟

到了20世纪90年代末，智能体范式已经成为人工智能中的主流正统理论：我们构建智能体，并赋予它们用户的偏好，智能体用理性的方式代表用户去做执行工作，使用如冯·诺依曼和摩根斯坦的预期效用理论之类的理性方式作为决策基础。智能体使用贝叶斯推断理性地管理它们对世界的信念库，通过贝叶斯网络或者其他方式捕捉对世界的理解。如果存在多个智能体，我们会借助博弈论来提供决策框架。这并非通用人工智能，也没有给出一条通往通用人工智能的光明大道，但到了20世纪90年代末，这些理念、工具和应用被接受的程度越来越高，致使人工智能界内部对人工智能本身的看法发生了重大变化。这是第一次，我们真实地感受到，自己并非在黑暗中拼命抓住所发现的任何东西。我们研究的领域已经有了明确而坚实的科学基础，从概率论到理性决策，这些都是经过时间检验、值得尊敬的技术。

这一时期的两项成就，标志着人工智能研究开始走向成熟。第一项已经成为全世界的头条新闻，而第二项，意义比第一项更为重大，但除了人工智能界内部人士以外，并不为人所知。

占据全世界头条新闻的突破来自IBM，　　1997年，IBM证明了名叫“深蓝”(DeepBlue)的人工智能系统能够在国际象棋比赛中击败俄罗斯大师加里·卡斯帕罗夫(Garry　Kasparov)。1996年2月，深蓝第一次从卡斯帕罗夫手里赢下棋局，不过他们一共下了六局棋，卡斯帕罗夫最终以4：2的总比分获得胜利。一年多后，升级的深蓝系统再次对阵卡斯帕罗夫，而这一次，人类的世界冠军被击败了。似乎当时的卡斯帕罗夫对此表示非常不满，并怀疑IBM公司存在作弊行为，但后来他也提到那是他国际象棋生涯中无法忘却的一次愉悦体验。从那以后，国际象棋无论从

任何意义上来说，都是一场已经没有悬念的游戏：人工智能可以战胜优秀的人类棋手。

深蓝的成功主要源于两个因素。第一是启发式搜索，由20世纪50年代跳棋程序的创造者亚瑟·塞缪尔提出，尽管这一技术经过了40多年的改进，其核心技术也不难理解。但第二个因素却颇具争议：深蓝是一台超级电脑，它依靠巨量的计算力来完成工作。有人批评说深蓝系统不是真正意义上的人工智能，只不过是靠野蛮的计算力来获胜的。但这种外行评论大大低估了人工智能技术的重要性，诚然，这些技术需要庞大的计算力支持，但在同样计算力的情况下，用简单粗暴的下棋方式，就如第二章里我们讨论过的解密汉诺塔步骤那种遍历的搜索方式，是根本不可行的。

可能你也听说过深蓝战胜卡斯帕罗夫的新闻，但我认为你不太可能听说当时的第二个重大成就，尽管它的影响可能更为深远。不知道你是否还记得，在第二章里，我们提到过某些计算问题本质上是复杂的——用技术语言来说，就是NP完全问题——人工智能领域许多亟待解决的问题都属于这一类。到了20世纪90年代初，人们开始清楚地认识到，解决NP完全问题的算法在进步，使得这个问题不像20年前一样成为人工智能不可逾越的障碍[61]。第一个被证实为NP完全问题的问题，就是SAT问题〔即satisfiability(可满足性)的缩写〕，这是一个检查简单逻辑表达式是否一致的问题——是否有任何方式表达它们都为真。SAT是所有NP完全问题中最基础的一个，你应该还记得，如果你能找到一个有效的方法解决某一个NP完全问题，那么你就自动找到了解决所有NP完全问题的方法。到了90年代末，“SAT求解器”，即解决SAT问题的程序已经足够强大，开始解决工业规模的问题了。在我写本书的时候，SAT求解器已经高效到在编程过程中可以随时调用了——它们已经成为我们解决NP完全问题的一个工具。不过这并不意味着NP完全问题已经被彻底解决，总会出现一些案例，让最好的SAT求解器都屈膝臣服。但我们不再害怕NP完全问题了，这是人工智能在过去40年里所取得的一项悄无声息的重大突破。

尽管人工智能领域刚刚重新获得了科学界的认可，但并不是所有人都为90年代末的主流人工智能感到高兴。2000年7月，我在波士顿参加会议，聆听了一名人工智能领域研究新星的演讲。当时有一位同行前辈坐在我旁边，他在黄金年代就从事了人工智能领域的研究，和麦卡锡、

明斯基是同一时期的人物。他很轻蔑地质问："这就是现在所谓的人工智能吗？魔法去哪儿了？"我能理解这句话的来历：现代的人工智能从业人员，需要的背景不再是哲学、认知科学和逻辑学，而是概率论、统计学和经济学。看上去就不那么具有……嗯……诗意了，对吧？但事实是，它成功了。

人工智能又一次改变了方向，但这一次，未来似乎更加稳定。如果我是在2000年写这本书的话，我肯定会预言，这个方向将是我从事的人工智能研究行业的未来。不会再有什么戏剧性的剧变发生了：我们将使用这些工具，并且，有了这些工具，我们的研究进展将是缓慢但稳定的。而那时候的我，并不知道戏剧性的新发展将再次震撼人工智能的核心。

第五章 深度突破

2014年1月，英国的科技行业发生了一件前所未有的事情：谷歌在这儿收购了一家小公司。虽然以硅谷的标准来看，这次收购很不起眼，但是在相对平稳的英国计算机行业来说，这是非同寻常的。这家被收购的公司名叫“深度思维”(DeepMind)，是一家新成立的公司，当时的员工不足25人，收购报价高达4亿英镑[62]。在外界看来，深度思维公司被收购时似乎没有任何产品、技术或者商业计划，它几乎不为人所知。甚至在它的专业领域——人工智能，亦如此。

谷歌斥巨资收购这家小小的人工智能公司上了新闻头条，全世界都想知道这家神秘的公司有些什么人，以及为什么谷歌认为一家名不见经传的小公司有着如此高的价值。

人工智能突然成了新闻热点——以及商业热点。全球对人工智能的兴趣激增，媒体也注意到了这股热潮，关于人工智能的报道频频见诸报端。各国政府也注意到了，并开始询问应该如何应对。一系列国家人工智能发展倡议很快接踵而至。科技公司都争先恐后投入这个领域，生怕被历史车轮甩落。随之而来的还有投资浪潮，虽然深度思维是人工智能领域最引人注目的收购，但也不乏其他案例。2015年，优步公司从卡内基-梅隆大学机器学习实验室揽获了至少40名研究人员。

在不到10年的时间里，人工智能突然从一潭计算机科学的死水变成最炙手可热、被炒作得最厉害的领域之一。人们对人工智能的态度发生突如其来的巨大变化，是由一项核心人工智能技术——机器学习的快速发展所推动的。机器学习是人工智能的一个分支领域，但在过去60年的绝大部分时间里，它一直在一条独立的道路上发展，正如我们即将在本章看到的，人工智能和这个突然爆发出魅力的分支学科之间，有时候关系会变得挺微妙。

在本章中，我们将了解21世纪机器学习革命是如何发生的，我们会从简要回顾机器学习开始，重点看一下机器学习的一种特殊方法——神经网络是如何在该领域占据主导地位的。就像人工智能本身的故事一样，神经网络的故事也充满了跌宕起伏：它曾经度过了两次寒冬，就在世纪之交的时候，许多人工智能研究人员还把神经网络视为一个毫无希

望的研究领域。但是，神经网络最终迎来了胜利，而推动其复兴发展的新思想是一种被称为深度学习的技术。深度学习正是深度思维公司的核心技术，我将告诉你深度思维的故事，以及深度思维构建的系统是如何引来全球赞誉的。但是，虽然深度学习是一门强大又重要的技术，但它并不是人工智能的终点。因此，正如对其他人工智能技术所做的那样，我们也将讨论它的局限性。

机器学习概述

机器学习的目标是让程序能够从给定的输入中计算出期望的输出，而不需要给出一个明确的方法。举个例子，机器学习的一个经典应用是文本识别，即获取手写文本并将其转录成文字。在此，给定的输入是手写文本的图片；期望的输出是手写文本表示的字符串。

任何一位邮政工作者都可以告诉你，文本识别挺难。每个人的字迹都不同，而且还有人笔迹潦草，有时候钢笔墨水会漏在纸上，有些书写的纸又脏又破。回顾一下第一章里面的图1，解读图片是一个未曾解决的问题。这不像玩棋盘类游戏，我们有理论上可以实现的步骤，只是需要启发式应用。我们并不知道图像识别的步骤是什么，所以需要一些特别的方式——这就让机器学习有了用武之地。

对于文本识别的机器学习程序，我们通常需要给它提供许多手写字符的范例来进行训练，每个范例都标有对应的实际字符，如图13所示。

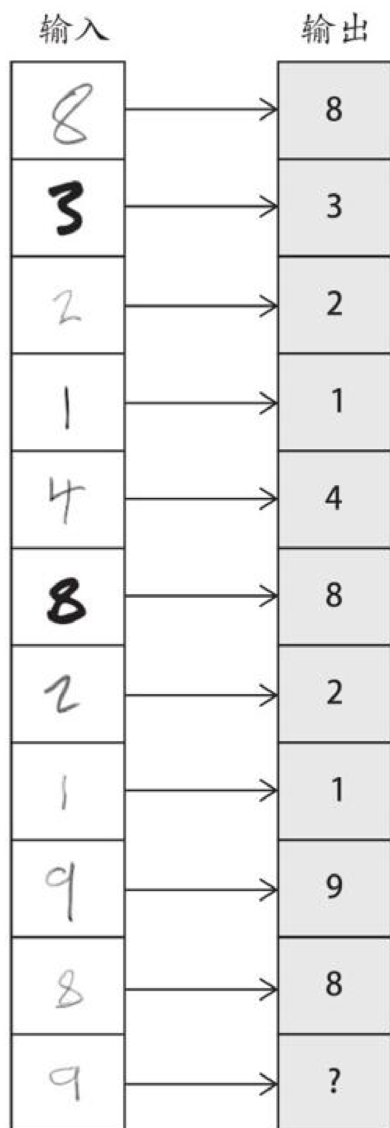


图13　程序训练数据

识别手写字符(本例中为数字)的机器学习程序训练数据，训练目标是这个程序能够自动识别手写数字。

刚才描述的这种机器学习类型被称为监督式学习，它有一个至关重要的要点：机器学习需要数据，大量的数据。事实上，正如我们即将看到的，提供精心制作的训练数据集合对机器学习的成功至关重要。

我们训练程序进行机器学习时，必须仔细设计训练数据集合。首先，我们通常只会使用一小部分可能的输入和输出来训练程序，在手写数字识别示例中，我们不可能向程序展示所有可能存在的手写字符，那根本不现实。再说了，如果我们可以向程序展示所有可能的输入集，那么就根本不需要机器来学习什么了：机器只需要记住每一个输入对应的输出就行了。无论何时对它进行输入，它只需要查找相应的输出即可——这不算机器学习。因此，一个程序必须只能使用可能存在的输入输出集合的一小部分进行训练，但是如果训练数据量太小，那么程序没有足够的信息来学会人们所期望的输入到输出的映射。

训练数据的另一个基本问题是特征提取。假设你在一家银行工作，银行需要一个机器学习程序来学习识别不良信贷风险。程序的训练数据是过去许多客户的记录，每个客户的记录上会标注他信贷记录是否良好。客户记录通常包括他们的姓名、出生日期、住址、年收入、交易记录、贷款记录和相应的还款信息等。这些信息在训练数据中被称为特征。不过，所有的特征在程序训练中都有意义吗？其中某些特征可能和该客户的信贷风险毫无关系。如果你事先不知道哪些特征和机器要学习的目标有关系，那么你可能试图将所有特征都放入训练数据中。但是，这样就会产生一个很严重的问题，被称为维度诅咒：训练数据包含的特征越多，你需要给程序提供的训练数据量就越大，程序学习的速度也就越慢。

最简单的应对方式就是只在训练数据中包含少量的特征，但这也会引起一些问题。一方面，你可能不小心忽略了程序正确学习所必需的特征，即确实标明客户信贷记录不良的特征，另一方面，如果你没有合理地选择特征，可能会在程序中引入偏差。例如，假设你给不良信贷风险评估程序训练数据里面导入的唯一特征是客户地址，那么很可能会导致程序在完成机器学习以后带有地域歧视。人工智能程序或许会变得有偏见，这种可能，以及它所引发的问题，我们将在后面详细探讨。

在另一种机器学习方式——强化学习中，我们不给程序任何明确的训练数据：它通过决策来进行实验，并且接收这些决策的反馈，以判断它们是好是坏。例如，强化学习被广泛应用于训练游戏程序。程序玩某个游戏，如果它赢了，就会得到正反馈，如果它输了，就会得到负反馈。不管正负，它得到的反馈都被称为奖励。程序将会在下一次玩游戏的时候考虑奖励的问题，如果它得到的是正面的奖励，那么下一次玩的

时候它更倾向使用同样的玩法，如果是负面的，那它就不太可能这样做。

强化学习的关键困难在于，许多情况下，奖励反馈可能需要很长的时间，这使得程序很难知道哪些行为是好的，哪些行为是坏的。假设强化学习的程序输了一场游戏，那么，究竟是游戏中的哪一步导致了失败呢？如果认为游戏中的每一步都是错误的，那肯定算总结过度。但我们怎么分辨究竟哪一步是错的？这就是信用分配问题[10]。我们在生活中也会遇见信用分配问题。如果你抽烟的话，很可能在未来收到与之有关的负面反馈，但是这种负面反馈通常会在你吸烟很久以后(通常是几十年)才会收到。这种延迟的反馈很难让你戒烟。如果吸烟者在吸烟以后立即就能收到负面反馈(以危及生命和健康的方式)，那么我认为，烟民数量一定会锐减。

到目前为止，我们还没提到程序是怎样进行学习的。机器学习作为一个学科领域，拥有同人工智能一样长的历史，也同样庞大。在过去的60年里，人们发展过各式各样的机器学习技术。不过近年来机器学习的成功源自一种特殊的技术：神经网络。其实，神经网络是人工智能中最古老的技术之一：1956年，约翰·麦卡锡在人工智能暑期学校里提出的最初建议就包括神经网络。但直到本世纪，它才再度引起了人们的广泛关注。

神经网络，顾名思义，灵感来自大脑内组成神经系统的神经细胞——神经元的微观结构。神经元是一种能够以简单的方式相互交流的细胞，自神经元发起的纤维突起，被称为轴突，与其他神经元进行连接，连接的“交叉点”被称为突触。一般来说，神经元通过突触连接来接收电化学信号，并且根据接收的信号，产生输出信号，然后由其他神经元通过突触连接接收。关键的是，神经元接收到的输入有着不同的权重：有些输入比其他的更重要，有些输入甚至可能抑制神经元，阻止它产生输出。在动物的神经系统中，神经元组成的网络是相互联系的：人脑大约有1000亿个神经元，人脑中的神经元通常有数千个连接。

因此，神经网络的构想，就是在机器学习的程序中引入类似的结构。毕竟，人类大脑已经充分证明了神经系统能够有效地学习。

感知器(神经网络1.0)

神经网络的研究起源于20世纪40年代美国研究人员沃伦·麦卡洛克(Warren McCulloch)和沃尔特·皮茨(Walter Pitts)，他们意识到神经元可以用电路建模，更具体地说，是用简单的逻辑电路，他们用这个想法建立了一个简单但非常通用的数学模型。到了50年代，弗兰克·罗森布拉特(Frank Rosenblatt)对这个模型进行了改进，创造出了感知器模型。感知器模型意义重大，因为它是第一个实际出现的神经网络模型，时至今日，它仍然有存在的意义。

图14展示了罗森布拉特的感知器模型，中间的方块代表神经元本身，左边指向方块的箭头代表神经元的输入(对应神经元的突触连接)，右边的箭头代表神经元的输出(对应轴突)。在感知器模型中，每一个输入都跟一个被称为权重的数字关联，在图14中，与输入1相关的权重为w_1，与输入2相关的权重为w_2，与输入3相关的权重为w_3。神经元的每一个输入都呈激活和未激活两种状态，如果一个输入被激活，它就会通过相应的权重“刺激”神经元。最后，每一个神经元都有一个触发阈值，由另一个数字表示(在图14中，触发阈值用T表示)。感知器的运作模式是神经元受到的刺激超过了触发阈值T，那么它就会“启动”，这就意味着它的输出被触发。换句话说，我们把激活的输入的权重加在一起，如果总权重超过阈值T，则神经元产生一个输出。

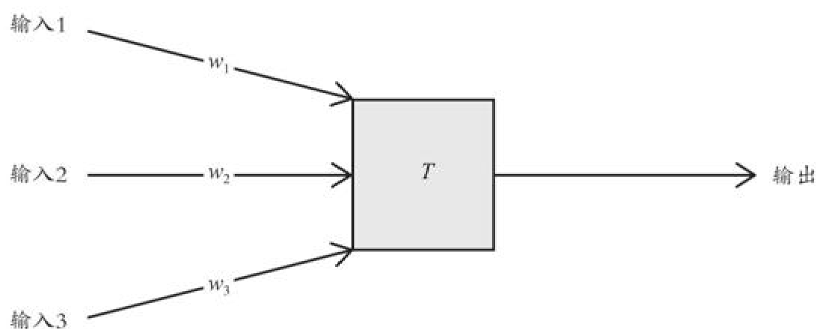


图14　罗森布拉特感知器模型中一个简单的神经元结构

具体来说，假设图14中神经元每个输入的权重都为1，阈值T为2。如果其中任意两个输入被激活，神经元就会启动输出。换言之，在这种情况下，超过半数的输入被激活，则神经元就会被触发。

我们再假设输入1的权重为2，而输入2和3的权重都为1，阈值T为2。在这种情况下，如果输入1激活，或者输入2和3共同激活，或者三个输入都激活，神经元就会被触发。

当然，真实存在的神经网络包括许多神经元，图15展示了由三个人工神经元组成的感知器。注意每个神经元都是完全独立运作的。此外，每个神经元能“看到”每一项输入——然而，对于不同的神经元，输入的权重可能不同。也就是说，输入1对于三个神经元分别有各自的权重值，可能它们并不相同。另外，每个神经元的触发阈值也可能并不相同(图15中分别为T_1， T_2和T_3)。所以，我们可以想象为三个神经元在各自计算不同的东西。

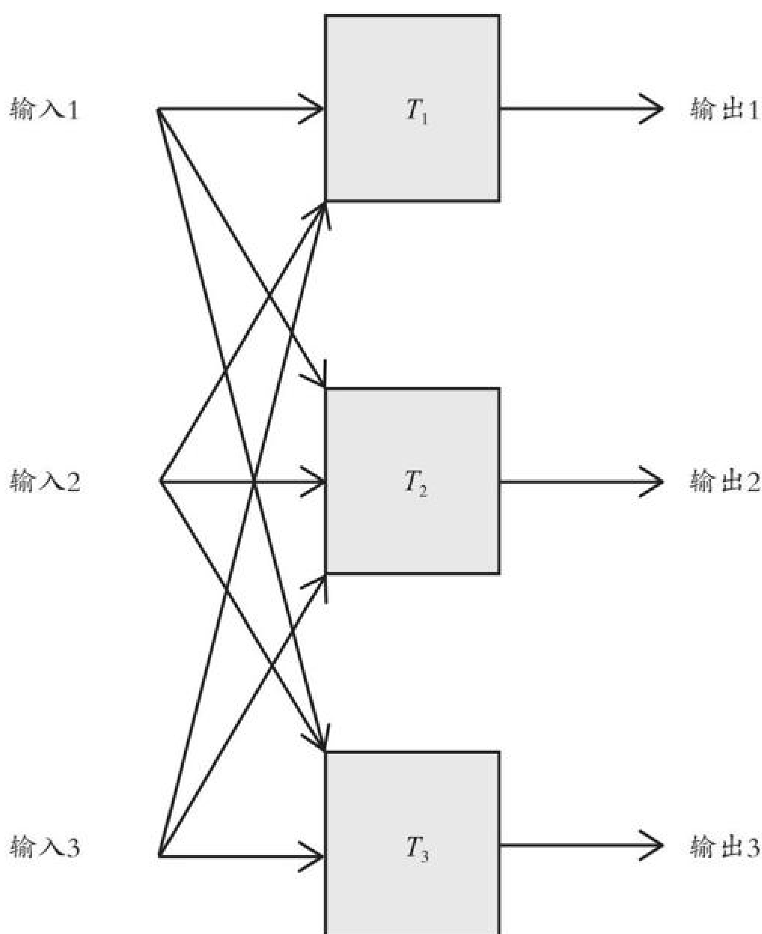


图15　由三个人工神经元组成的单层感知器

然而，图15所展示的感知器并没有反映出大脑高度互联的结构，一个神经元的输出会反馈给其他许多神经元。为了更清楚地反映人脑结构的复杂性，人工神经网络通常是分层组织的，如图16所示，即多层感知器结构。图16的感知器由9个神经元组成，分为3层，每层3个神经元。每一层的每个神经元都接收上一层神经元的输入。

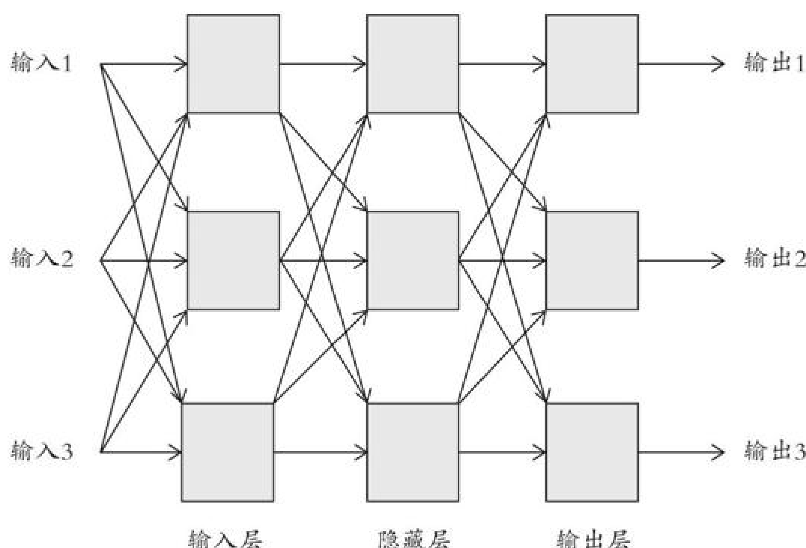


图16　由9个神经元组成的3层感知器

需要注意的是，即使在这个非常简单的感知器中，事情也开始变得复杂了：我们的神经元之间已经有27个连接了，每个连接都有对应的权重，9个神经元都有自己的触发阈值。虽然麦卡洛克和皮茨就在模型中设想了多层神经网络结构，但在罗森布拉特的时代，人们主要关注单层网络，原因很简单：没有人知道如何训练具有多个层面的神经网络。

每个连接所对应的权重值对于神经网络的运行至关重要，事实上，这就是神经网络分解下来的全部内容：一堆数字列表。对于任何一个大小合理的神经网络来说，这个数字列表的长度都相当可观。因此，训练一个神经网络需要用某种方式找到适当的权重值。通常的寻找方式是在每次训练以后调整权重值，试图让网络产生正确的输入到输出的映射。

罗森布拉特试验了几种不同的技术，并为一个简单的感知器模型找到了一个被他称为纠错程序的技术。

现在我们知道罗森布拉特的方法肯定是有效的，它可以正确地训练一个网络。但是在当时，存在一个强烈的异议。1969年，马文·明斯基和西摩·帕普特(Seymour Papert)出版了一本名叫《感知器》的书[63]，书中指出单层感知器网络有着非常大的局限性。事实上，如图15所示的单层感知器确实如此，它们甚至连许多输入和输出之间的简单关系都学不会。但当时吸引大多数读者注意力的，是明斯基和帕普特的研究表明，感知器模型不能学习一个很简单的逻辑概念——异或(XOR)[11]。举一个例子，假设你的网络只有两个输出，当其中一个输出被激活时，异或函数应该产生一个输出(但当两个输入同时被激活的时候，则不会产生输出)。要证明单层感知器无法表示异或状态很容易，感兴趣的读者可以在附录D中找到更多信息。

明斯基和帕普特的书似乎给出了相当全面的结论，不过时至今日仍然存有争议。该理论结果证明了某些类别的感知器在基础结构层面具有严重局限性，这似乎就意味着基于感知器的通用模型存在局限性。而如图16所示的多层感知器并不受这些限制：从精准的数学定义来说，可以证明多层感知器完全能够普遍适用。然而，在当时，没人知道该如何训练一个具有多层感知器的网络：它只是一个理论上可能出现的网络结构，在现实中无法构造。20年后，随着科学的发展，它才从理论走向实践。

我很怀疑，当年对感知器太过激进的宣传间接导致了对它的负面结论下得如此武断。比如，1958年《纽约时报》上某篇文章兴奋地报道[64]：

> 美国海军今天公布了一个电子计算机雏形，人们期望它能够行走、说话、视物、书写、自我复制，并且意识到自己的存在。

对于神经网络研究衰落的确切原因，我们可以展开各种辩论，但不管是什么，到了20世纪60年代末，神经网络研究急剧衰落。人们转而支持麦卡锡、明斯基和西蒙倡导的符号人工智能的方式(讽刺的是，神经网络研究的衰落仅仅发生在人工智能寒冬——我们在第二章里提到过——出现的前几年)。1971年，罗森布拉特死于一次航海事故，使得神经网络研究领域失去了一员主力大将。如果他能活下来，人工智能的历史也

许会有所不同。总之，在他死后，神经网络的研究被搁置了十多年。

连接主义(神经网络2.0)

神经网络研究领域一直处于休眠状态，直到20世纪80年代才开始复苏。一本分上下两册出版的书籍《并行分布式处理》[65]，预示着神经网络研究领域的复兴。并行和分布式处理(简称PDP)是计算机科学研究的一个主流领域，它主要研究如何建立能够并行运算的计算机系统。乍一看，这本书跟人工智能或者神经网络毫无关联，而且我想某些看了书名就买书的人，在发现本书的内容和神经网络有关的时候，会感到无比困惑。或许作者选择这个标题就是为了跟之前的神经网络研究撇清关系吧。

从某种意义上来说，新兴研究的最重要部分也没那么新颖：它主要研究多层神经网络，可以轻易克服明斯基和帕普特所断定的简单感知器系统的局限性。不过跟之前的研究仍然有一点关键的区别。以前关于感知器的研究主要集中在单层网络上，因为当时没有人知道如何“训练”多层神经网络，也不知道如何找出神经元之间连接的权重值。PDP以一种被称为反向传播的算法为这个问题提供了解决方案，这或许是神经网络领域中最重要的一门技术。

就如科学研究中经常发生的情况一样，反向传播似乎在过去的几年里被发明和重新发明过很多次，但是PDP研究人员引入的特定方法最终确定了它的地位[66]。

如果要完整地解释反向传播算法，我们必须引入本科水准的微积分知识，这远远超出了本书设定的范围。不过反向传播算法的基本思想很简单，它的工作原理是收取神经网络出错的反馈，这里的错误是在网络的输出层的输出(比如网络输入了一张猫的图片，而输出层将其归类为一条狗)。反向传播算法将错误从输出端向输入端逐层逆向修正(算法也是因此而得名的)。它首先计算误差值(即输出的数据和期望得到数据之间的差值)，在给定输入和输出的情况下，误差是一个和权重有关的函数，需要通过修正权重值使得误差值达到极小(即尽量减少误差)。根据误差值能够得到等值线图，在等值线图上体现为最陡的下降路线，即为从当前的误差到我们期望的最小误差的方法。这个通过调整权重值来减少误差值，最终接近极小误差(即输出结果尽量接近期望输出)的过程被称为梯度下降。然后，调整完最后一层权重以后，逐级往前调整，以此类

推。

PDP还提供了比感知器更适用的神经元模型。感知器模型本质上还是二进制的计算单元(状态为开或者关)，而PDP的神经元模型更具备通用性。

反向传播算法的发展和PDP研究界引入的其他创新，使得神经网络具备广泛应用的可行性，这远远超出了20年前感知器模型的简单演示，人们对神经网络的发展兴趣倍增。但事实证明，PDP的泡沫也没有持续太久。到了90年代中期，神经网络研究再次失宠。事后看来，神经网络车轮从PDP研究这辆马车上脱落的原因，并非研究基础有固有缺陷这类硬伤，而是源于一个平淡无奇的理由：当时的计算力不够强大，无法承载新技术。并且，PDP的进步似乎十分缓慢，而机器学习的其他领域又在飞速发展，因此，机器学习的主流热点，又一次从神经模型上转移开了。

深度学习(神经网络3.0)

我想起2000年前后参与的一个学术人工智能特派专家组的经历。小组的一位成员试图说服我们拒绝任何从事神经网络研究的申请人。“这是一个富有影响力、充满机会的领域，”他辩称，“我们为什么要雇用一个研究夕阳产业的人呢？”不过我们忽视了他的意见。但公正地说，在2000年，你必须具备非凡的远见卓识才能预测到神经网络即将再次复兴。到了2006年前后，一场复苏确实开始了，它引起了人工智能史上规模最大、宣传最广的爆发。

推动第三次神经网络研究浪潮的关键技术被称为深度学习[67]。我倒是很乐意告诉你深度学习可以用某个单一的关键理念描述出来，可惜，事实上，这个术语指代的是一系列相关思想的合集。深度学习至少可以从三个不同的方面解读。

其中最重要的，顾名思义，就是网络要具备“深度”，即多层结构。每一层可以在不同的抽象层面上处理一个问题——靠近输入层的层面处理数据中比较低级的概念(例如图片的边缘之类)，而越是到了深层网络，就处理越为抽象的概念。

深度学习不仅仅体现在“深度”上，还能够享受神经元数量剧增的益处。一个典型的1990年的神经网络可能只有大约100个神经元(如果你没

忘的话，人类的大脑大约有1000亿个神经元)。这样的网络在处理具体问题上显然十分具有局限性。到了2016年，先进的神经网络已经拥有大约100万个神经元了(这个数量和蜜蜂的大脑大致相同)[68]。

最后，深度学习使用的深层次网络中，神经元本身的连接数量也十分可观。在20世纪80年代出现的高度连接神经网络中，每个神经元可能与其他神经元产生150个连接。到了撰写本书的时候，最先进的神经网络中的神经元，已经和猫的大脑神经元连接数相当了。而人类的神经元平均拥有10 000个连接。

现在，深度神经网络拥有更多的网络层次结构、更多的神经元以及每个神经元拥有更多的连接，为了训练这样的网络学习，就需要比反向传播算法更先进的技术。杰夫·辛顿(Geoff Hinton)于2006年提出了这一观点，他是一位英国出生的加拿大研究员，比任何人都认同深度学习的改革。不管怎么说，辛顿是个了不起的人，他也是20世纪80年代PDP运动的领导人之一，同样也是反向传播算法的创始人之一。我个人认为他最了不起的一点在于，当PDP研究失宠后，辛顿并没有灰心丧气，而是坚持下来，并以深度学习的形式将神经网络带入另一个辉煌，他也因此受到了国际社会的赞誉。(很凑巧，辛顿正好是乔治·布尔的曾孙，我们在第三章里提到过布尔，他是现代逻辑的奠基人之一。不过，辛顿声称，或许这是他和逻辑派传统人工智能唯一的关联。)

更深的网络层级、更庞大的神经元结构、更广泛的神经元连接，是神经网络深度学习模式成功的一个关键因素。而辛顿和其他人在关于训练神经网络方面提供的新技术是另一个关键因素。但深度学习真正获得成功，还需要另外两个因素：数据和计算能力。

数据对机器学习的重要性可以用ImageNet项目的故事来说明[69]。ImageNet来自华裔研究员李飞飞的创意。1976年她出生于北京，80年代随父母移居美国，学习物理和电气工程。2009年，她进入了斯坦福大学，并在2013年至2018年间带领斯坦福大学人工智能实验室。李飞飞认为，机器学习需要大型的、维护良好的数据集，这可以为新系统的训练、测试和比较提供一个通用的基线，也将使整个深度学习研究界受益匪浅。因此，她启动了ImageNet项目。

ImageNet是一个大型的在线图像档案库，在撰写本书时，已经拥有大约1400万张图片。ImageNet的图片仅仅是照片而已，你可以下载为

普通的数码格式，比如JPEG。不过，最重要的是，这些图片被详细分为22 000种不同的类别，使用一个名为“词汇网”[70]的在线语义词库标注。词汇网的单词被仔细分类过，例如可以识别具有相同或者相反含义的词汇等等。现在查看ImageNet的图片，我们可以看到它包含1032张标记为“火山口”的图片，122张标记为“飞盘”的图片，诸如此类。我们需要了解的一个重点是，数据库中特定类别的图像并非人工分类的，也不是因为看上去很相似所以列入分类——恰恰相反，举个例子，飞盘类的图片唯一的共同点是它包含飞盘。其中某些图像是一个人朝另一个人扔飞盘，也有图像是静止在桌面上的飞盘，没有任何人影。每张图片都不一样——除了它们都包含飞盘这个要素。

2012年是技术图像分类发展的最佳时机，当时杰夫·辛顿和他的两位同事亚历克斯·克里泽夫斯基(Alex Krizhevsky)、伊利亚·苏茨科弗(Ilya Sutskever)一起展示了名为**AlexNet**的神经网络系统，它在国际图像识别比赛中有着亮眼的表现[71]。

使深度学习发挥作用的最后一个要素是计算机的处理能力。训练一个深度神经网络需要大量的计算机处理时间，训练本身要做的工作并不太复杂，但是数量庞大。21世纪初开始流行的一种新型计算机处理器被证明是计算繁重任务的理想选择。图形处理单元(GPU)最初是为了处理计算机图形问题而开发的，例如为电脑游戏中提供高质量的动画。但这些芯片被证明是训练深度神经网络的完美工具。现在，每一个名副其实的深度学习实验室里都有GPU群——然而，不管它们拥有多少GPU，实验室的工作人员都会抱怨还不够。

毋庸置疑，深度学习和神经网络取得了成功，但它们也存在一些众所周知的缺点。

首先，它们所体现的智慧是不透明的。神经网络所获取的知识体现在神经元之间相互连接的权重值上，到目前为止，我们还没有办法解析这些知识。一个深度学习的程序可以告诉你在X光扫描图片中哪里有肿瘤，但它无法证明它的诊断是正确无误的。一个拒绝为客户提供银行贷款的深度学习程序无法告诉你它拒绝客户的原因。在第三章中，我们看到类似MYCIN这样的专家系统能够对系统结论做出粗略解释——专家系统得出结论的推理依据是可以追溯的，但神经网络无法做到这一点。目前有许多研究人员正在致力于解决这个问题，但是，到现在为止，我们还不知道如何解释和表达神经网络所包含的知识。

另一个关键问题是神经网络的稳定性，这是个不易察觉但非常重要的问题。例如，如果对图像进行细微的修正，对人类而言，这种修正完全不会影响图像识别，但会导致神经网络错误地将其分类，如图17所示[72]。图a是熊猫的原始图像，图b是经过修改的。我想你会认为这两张图没什么差别，而且你肯定会认同它们都是熊猫图片这个结论。可是神经网络能够正确地将图a分类为熊猫，但对图b，它则错误地将其分类为长臂猿。为了解决这类问题而进行的研究被称为对抗性机器学习——这个术语源自一个观点，即有对手故意通过修改图片参数的方式来试图蒙骗程序。



图17　熊猫还是长臂猿？

神经网络可以正确地将图像a分类为熊猫。然而，以人类无法察觉的方式调整过图像以后，同一个神经网络会错误地将图像b分类为长臂猿。

图像分类程序认错了动物，倒不是什么要紧的事情，但对抗性机器学习已经昭示了一些令人惶恐的案例。例如，事实证明，同样的改图方式可以影响程序对路标的识别，虽然人类肉眼看来没什么区别，但是在无人驾驶汽车中，神经网络就可能误读路标。所以，我们要在敏感的应用程序中使用深度学习算法，就需要详细了解这方面的问题。

深度思维

在本章最开始提到的深度思维公司的故事完美地代表了深度学习的兴起，这家公司由人工智能研究人员兼电脑游戏爱好者德米斯·哈萨比斯

(Demis Hassabis)和他的校友，企业家穆斯塔法·苏莱曼(Mustafa Suleyman)于2010年创立，公司还有哈萨比斯在伦敦大学学院工作时认识的计算神经科学家谢恩·莱格(Shane Legg)加盟。

正如我们所知的，谷歌在2014年收购了深度思维公司，我还记得在媒体上看到这则新闻时，惊讶于深度思维是一家人工智能公司。显然，在收购的时候人工智能成为新的热点，这不足为奇，我惊讶的是在英国竟然有一家我没有听说过的人工智能公司，价值4亿英镑，这真的让我困惑不已。和别人一样，我立马访问了深度思维的网站，但坦白地说，这并没有为我解惑。公司的技术、产品和服务都没有任何详细的说明。然而，它倒是提供了一个有趣的话题：深度思维公司公开宣称其任务是解决智能问题。我已经提到过，人工智能在过去60年里命运多舛，这让我对任何雄心勃勃预测人工智能进步的消息保持警惕：你可以想象，看到一家刚刚被科技巨头收购的小公司竟然发表如此大胆的声明，让我多么吃惊。

不过网站上没有提供更多细节，我当时也没能找到更了解它的人。人工智能界的同行大多对这个声明持怀疑态度，可能也带有一点职业嫉妒的色彩。直到2014年末，我碰巧遇见了一位同事南多·德·雷弗斯塔(Nando de Freitas)，才打听到更多关于深度思维的消息。南多是深度学习领域世界级先驱之一，当时他是牛津大学的教授，是我的同事(后来他离开学校，去了深度思维工作)。他正好和学生们参加一个研讨会，腋下夹着一堆科学论文。显然他在为某件事情兴奋着，他告诉我："伦敦有人训练出了一个程序，可以从头开始玩雅达利游戏。"

我不得不说，这有什么好兴奋的？可以玩电子游戏的程序又不是什么新鲜事。这种程度的挑战我们可能会布置给本科生作为毕业项目——我就是这么轻蔑地告诉南多的。他很耐心地给我详细解释，我们确实已经进入了人工智能的新时代。

南多提到的雅达利游戏系统基于早期的雅达利2600系列游戏机，那是1980年前后的产品，是最早获得成功的视频电子游戏平台之一：它支持210×160像素网格的大分辨率视频，支持128位颜色。用户通过一个带单独按钮的操作杆进行操作，游戏机使用插卡式游戏卡带，深度思维用的游戏卡带一共有49个游戏。他们对人工智能的描述如下[73]：

> 我们的目标是构建一个单一的神经网络游戏智能体，它能够学习玩尽可能多的游

戏。神经网络没有提供任何特定游戏的信息或者额外加入的视觉要素，也不了解游戏机的内部状态。它只能从显示屏所反馈的东西进行学习(即游戏分数)，以及弄清楚到底该怎么操作游戏——就像人类玩家一样。

为了理解深度思维成就的意义，了解他们的程序到底做了什么以及没做什么，这一点非常重要。或许最重要的一点是，程序对自己正在玩的游戏一无所知。如果我们试图用第三章中提到的基于知识的人工智能来构建一个雅达利游戏程序，首先会考虑从雅达利游戏机中提取的专家系统知识，并尝试使用规则或其他一些表示知识的方案对其进行编码(如果你想试试的话，祝你好运)。但是深度思维的程序根本没有任何关于游戏的知识，程序得到的唯一信息是出现在游戏机屏幕上的图像(以210×160彩色像素网格的形式)和游戏的当前分数。仅仅如此，这个程序压根没有其他的信息可以参考。这里要特别提出来讲一下，程序没有得到诸如“对象A在位置(x， y)上”之类的信息——任何类似的信息都需要程序从原始的视频数据中自己提取。

程序运行的结果简直令人惊讶[74]，程序通过强化学习自学玩游戏：反复玩同一个游戏，在每个游戏中进行实验并获得反馈，并学习哪些行为会得到奖励，而哪些不会。雅达利游戏程序学会了游戏卡带中的29个游戏，表现出高于人类玩家平均水准的能力。在某些游戏里面，它甚至达到了超人的水准。

有一种特别的游戏吸引了很多人的注意。这款名为“打砖块”的游戏是20世纪70年代最早出现的电子游戏之一：由玩家控制“球板”，将“球”弹到由彩色“砖块”组成的墙上。每当球碰到砖块的时候，砖块就会被摧毁，玩家的目的是摧毁整面砖墙。图18显示了程序在学习玩打砖块的早期阶段(在它玩了大约100次游戏之后)，在这个阶段，它经常漏球。



图18　程序学习玩打砖块游戏的早期阶段

深度思维的雅达利游戏程序在学习打砖块的初期漏接了球，球在这里用一个竖直的小矩形表示，水平的平板状矩形表示玩家的接球板。

但是经过几百轮的训练以后，程序就成了这个游戏的专家：它再也没有漏接过一个球。然后发生了一件不同寻常的事情：程序了解到，最有效率得高分的方式是在砖墙的一侧“钻”一个洞，让球打进去，这样球就会在砖墙和顶部屏障之间快速反弹，迅速消灭砖块，而玩家可以不用额外操作什么(见图19)。深度思维公司的工程师并没有预料到这种行为：它是由程序自主学习的。这个游戏的视频很容易在网上找到：我在自己的讲座中用过十几次。每次我给观众播放这段视频时，都能听见惊讶的抽气声，因为观众明白程序在游戏中学到了什么。



图19　程序经过训练后玩打砖块游戏

最终，程序学会了怎么迅速取得高分，即让球在砖墙一侧“钻”一个洞，这样球就会在砖墙上快速反弹。没有人教程序这么做，这种行为让程序开发者都大吃一惊。

我得反复强调一点：深度思维的程序员并没有编写一个程序来玩雅达利游戏：这并不难。他们所做的是写一个程序，让它学习如何比人类更会玩全部49个雅达利游戏中的29个。程序接收到的唯一输入就是屏上显示的东西，以及分数。

此前已经提到过，雅达利游戏程序使用的是强化学习的方式，通过神经网络来实现，它使用的神经网络具有三个隐藏层。神经网络的输入经过预处理，将图像从原始的210×160像素的彩色格式缩减为84×84像素，并用灰度代替了彩色。程序从可用的输入中提取出样本，由每四幅游戏屏幕图像拼合组成，而不是单独的每一幅图像。神经网络使用经典的深度学习技术(随机梯度下降)进行训练。

这个程序当然不会是完美的，在某些游戏中，它的表现相当糟糕，研究一下为什么会出现这种情况也挺有意思。在程序玩得特别糟糕的游戏中，有一款叫作“蒙特祖玛的复仇”，它的难点在于奖励非常稀少：玩

家在获得奖励之前必须执行一系列复杂的任务(这一点与打砖块这种游戏不同，在打砖块游戏中奖励反馈或多或少都是即时的)。通俗地说，如果奖励反馈在相关行动执行后很长时间才出现，就会给强化学习带来困难：这就是前文我们讨论过的信用分配问题，即你可能不清楚是哪些行为导致了奖励的发生。

如果雅达利游戏程序是深度思维团队唯一完成的东西，那也足够让他们在人工智能的历史上留下令人尊重、浓墨重彩的一笔，但是，该团队随后又取得了一系列惊人的成就。

其中最著名的是AlphaGo，在撰写本书的时候，它可能仍然是迄今为止最著名的人工智能系统。AlphaGo的功能是玩一种源自中国的古老棋类游戏：围棋。

围棋是人工智能挑战一个引人关注的目标。一方面，围棋的规则非常简单，比国际象棋简单得多。另一方面，在2015年，围棋程序的水准远远低于人类专业棋手。那么，为什么围棋对人工智能而言这么难？答案很简单，因为围棋的计算量太庞大了。围棋棋盘是19×19的格子，总共有361个位置可以落子。而国际象棋棋盘的格子是8×8，只有64个位置可以放置棋子。正如我们在第二章中所提到的，围棋的分支因子(即棋手在游戏中每一步的平均移动可能性)约为250，而国际象棋的分支因子约为35。换言之，在棋盘大小和分支因子方面，围棋的数据量比国际象棋庞大得多。另外，一盘围棋对弈可以持续很长时间，一场比赛中走150步是很常见的。

对人类而言，围棋是公认最难的棋类，因为计算规模太大：思考一个如此大小的棋盘已经达到甚至超过人类玩家所能管理的极限。这导致围棋中指定明确的战术非常困难。对于机器而言，这也是问题所在。棋盘规模和分支因子让简单粗暴的搜索方式毫无用武之地——我们得考虑别的方法。

AlphaGo使用了两个神经网络：价值网络只负责评估给定的棋盘位置的优劣程度，而策略网络则根据当前棋盘的状况评估下一步棋该放在何处[75]。策略网络包含13层，首先使用监督式学习进行训练，训练的数据则是人类的专业棋手下棋的棋谱。然后进行自我对战的强化学习。最后，这两个网络被嵌入一个复杂的蒙特卡罗树这一搜索技术中。

在这套系统公布之前，深度思维邀请了欧洲围棋冠军樊麾与AlphaGo比赛：最终AlphaGo以5：0获胜。这是围棋程序第一次在全场比赛中战胜人类专业棋手。不久之后，深度思维宣布AlphaGo将于2016年3月在韩国首尔与世界围棋冠军李世石进行五场比赛。

人工智能界因此兴奋不已，相关研究人员——包括我自己——也很期待看到比赛结果(当时我们猜测AlphaGo大概会取得一到两场胜利，但李世石会决定性地赢得整场比赛)。谁也没有料到围绕这场比赛爆发出空前的宣传热浪，这项赛事成为全世界的头条新闻，比赛的故事甚至都被拍成了电影[76]。

这场比赛中，AlphaGo以4：1的成绩击败了李世石：李世石输掉了前三局，在第四局中扳回一城，但输掉了第五局。大家都说，李世石一开场就输了。他很惊讶——本来他以为会轻松取得胜利。而且不止一个人开玩笑说，AlphaGo故意输掉第四局，多少给李世石留点面子。

在比赛的很多节点上，人类评论员指出AlphaGo的举动奇怪。很明显，这种落子“不是人类会做的选择”。当然，当试图分析AlphaGo如何下棋的时候，我们是从一个非常人性化的角度来分析的，我们本能地寻找下围棋时人类的动机和策略——我们将AlphaGo人性化了。试图用这种方式理解AlphaGo是没有意义的：它只是一个程序，它的存在只为了一个目的——赢得围棋比赛。我们想把动机、推理和策略归因于程序，但无法做到，AlphaGo的卓越能力是通过其神经网络的权重来体现的。这些神经网络不过是一串很长的数字列表，我们无法提取或合理化它们所包含的专业知识。AlphaGo也无法告诉我们它如此落子的原因，而这正是深度学习需要解决的关键问题之一。

AlphaGo被吹捧为深度学习和大数据型人工智能的胜利，从事实来看，它倒也实至名归。不过撇开表象深入挖掘，你会发现AlphaGo中最能体现智慧的工程都源自经典的人工智能搜索。我们在第二章提到的于20世纪50年代开发了跳棋学习程序的亚瑟·塞缪尔，他在理解AlphaGo使用的搜索技术时不会有任何困难：从他的跳棋程序，到现代最引人注目的人工智能系统，都遵循着同一条发展路径。

两个里程碑式的成就，对大多数人而言应该已经足够了。但是，仅仅18个月后，深度思维再次出现在头条新闻中。这次是一个比AlphaGo更厉害的人工智能，被称为AlphaGo Zero。它的非凡之处在于它是从零

开始学习下围棋的，没有学习任何人类棋手的棋谱。在没有人工数据干预的情况下，它达到了超越人类棋手的水平，而这一切，只是通过它自己和自己下围棋来实现的[77]。公平地说，它必须自己下过很多次才能达到超人的水准，但不管怎么说，这都是一个惊人的成就。它的后续版本名叫AlphaZero，进一步推广到玩包括国际象棋的其他棋类游戏。在结束9个小时的自我学习以后，AlphaZero能够在和鳕鱼系统[12]对战中连续击败对方，最少也能保持平局——鳕鱼系统是世界领先的国际象棋程序之一。来自国际象棋编程领域的退役专家纷纷表示他们极其惊讶。AlphaZero自己和自己下了9个小时国际象棋，就能够自学成为世界级的国际象棋手？这种想法简直令人难以置信。而真正令人兴奋的是这种方法的普适性：AlphaGo尽管在围棋方面表现优秀，但它只能下围棋，还必须事先学习许多人类专业棋手的棋谱。而AlphaZero似乎可以自学成才，并且适用于多种不同类型的棋类游戏。

当然，我们得谨慎地下结论，不能过度解读。首先，尽管AlphaZero体现了令人印象深刻的通用性(它在棋类游戏专业的通用性方面超过了此前任何一个人工智能系统)，但它本身并不代表迈向通用人工智能的重大进步。它甚至没有我们人类所普遍拥有的智能，在下棋方面它很专业，但它不能跟人交流，不能讲笑话，也不会煎鸡蛋、骑自行车或者系鞋带。它的卓越能力其实有着相当的局限性。当然，棋类游戏是相当抽象的——它们与现实世界相去甚远，正如罗德尼·布鲁克斯很快将要提醒我们的那样。

但是，尽管还有许多问题和麻烦，我相信一个简单的事实：深度思维的工作，从他们的雅达利游戏机到AlphaZero，代表了人工智能领域一系列非凡的突破性成就。在实现这一切的过程中，他们成功地让数以百万计的人们美梦成真。

迈向通用人工智能？

深度学习已经被证明成就非凡，它使我们有能力构建一些在几年前无法想象的人工智能程序。尽管这些程序赢得了辉煌胜利，但它们并不是推动人工智能朝着宏伟梦想前进的魔法。接下来，为了解释这个问题，我们来看一下两个现在广泛使用了深度学习技术的应用：图像标注和自动翻译。

在图像标注问题中，我们希望计算机能够获取图像并对其进行文本

描述。在某种程度上具备这项功能的系统已经得到广泛应用：我的苹果Mac软件在更新照片管理应用程序以后，能够正确将我的照片分为“海滩场景”“派对”等等。在撰写本文的时候，还有好几个通常由国际研究机构运营的网站存在，你可以将照片上传到网站，它会尝试为照片做出标识。为了更好地理解图像标注技术的局限性，进而理解深度学习的局限性，我将一张家庭照片上传到一个网站中(本例中，我使用的是微软的标注机器人)[78]，照片如图20所示。

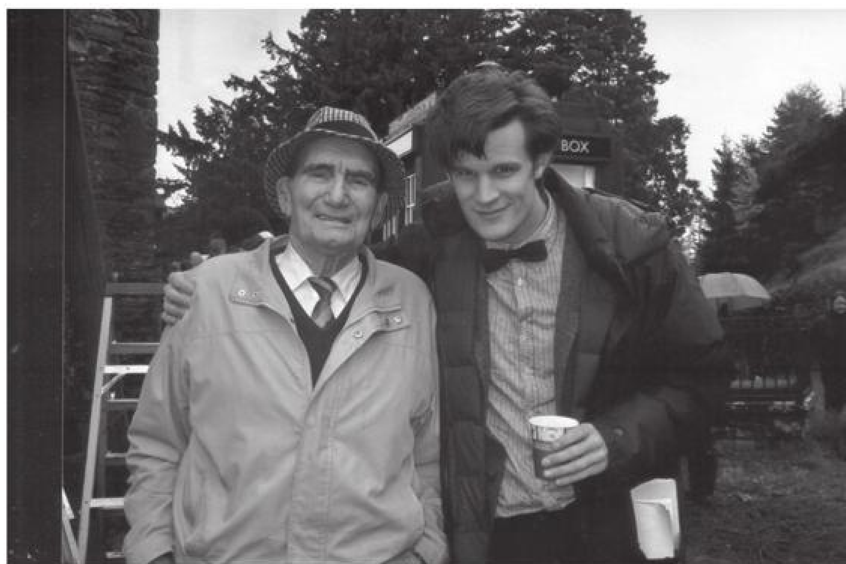


图20　这张照片的内容是什么呢？

在我们得知标注机器人的回应之前，先请你看看这张照片。如果你是个英国人，或者是科幻小说迷，那么你可能会认出照片中右边这位先生是马特·史密斯(Matt Smith)，他在2010年至2013年的BBC电视节目中扮演神秘博士(左边那位就别去猜了，那是我已故的岳父)。

标注机器人对照片的回应如下：

> 我想这是马特·史密斯以站姿拍照，他们看上去似乎很:-) :-)

标注机器人正确地识别了照片中的关键元素，并在某种程度上识别了照片背景(站姿，拍照，微笑)，然而这种正确识别容易让我们误以为标注机器人正在做一些它肯定做不到的事情：理解。为了说明这一点，

请考虑系统是如何识别马特·史密斯的，正如我们之前所提到的，像标注机器人这样的机器学习系统是通过给它大量的数据作为训练样本训练出来的。每个训练数据都由图片和对应文字组成，最终，在识别了大量马特·史密斯的照片以及对应的文本(即“马特·史密斯”的人名)之后，当他出现在照片里，系统就能正确识别出来，并生成文本“马特·史密斯”。几十年的努力研究毕竟是有用的。

但标注机器人并没有真正“认出”马特·史密斯，为了理解这一点，假设我让你看这张照片，你可能会给我这样的回应：

> 这不是马特·史密斯吗？演神秘博士那位演员，他搂着一个老人站着，这个老人我不认识。他俩都在笑。马特打扮成神秘博士的样子，可能是在拍摄现场吧。他口袋里有卷起来的纸，大概是剧本。马特手里拿着纸杯，或许是在拍摄现场休息。背后的蓝色盒子，那不是塔迪斯吗？神秘博士的太空船时间机器，博士乘坐它四处旅行。他们是在户外拍摄这张照片的，所以很可能就是在摄影现场，附近可能会有摄制组、摄像机和灯光。

标注机器人无法做到这些，虽然它能够识别马特·史密斯，但无法正确理解此处的文本“马特·史密斯”意味着什么。它也无法利用这些知识来解释图片中正在发生的事情。缺乏理解，这就是此处的要点。

当你看到马特·史密斯打扮成神秘博士的照片，就可能联想到一系列的东西，而不仅仅是简单地识别出图片中的人物和解释图片本身。如果你是一个“神秘博士”的粉丝，甚至还有可能深情地回忆起你最喜欢的由他出演的电视剧的某一集(我选择《等待的女孩》，大家同意吗？)。你可能还会记起跟父母或者孩子一起看马特·史密斯主演的《神秘博士》时的场景，里面的怪物让你吓了一跳，等等；或者它会让你联想起一个摄影棚，或者摄制组什么的。

因此，你对这幅图的理解是基于你在这个世界上作为一个人类存在的经历。这样的理解对于标注机器人而言是不可能实现的，因为它没有这个基础(当然，它也并不打算拥有)。标注机器人完全脱离了现实世界，正如罗德尼·布鲁克斯提醒我们的那样：智慧是具体化的。我强调，这个观点并非认为人工智能系统无法做到理解，而是说理解并不是仅仅将某个输入(本例中指包含马特·史密斯的照片)映射到某个输出(本例中指文本“马特·史密斯”)。这种映射的能力可能是理解的一部分，但绝不是全部。

将一种语言自动翻译成另一种语言，是过去十年中因为深度学习技术而快速进步的另一个领域。来看看自动翻译工具能做到什么，又不能

做到什么，有助于我们理解深度学习的局限性。谷歌翻译可能算是最著名的自动翻译系统了[79]，作为一个产品，它最初于2006年推出，最新版本的谷歌翻译使用深度学习和神经网络，这个系统是通过给它大量的翻译文本训练出来的。

让我们看看，2019年版本的谷歌翻译遇见不合理的难题时会怎么处理。我们让谷歌翻译法国作家马塞尔·普鲁斯特(Marcel Proust)在20世纪早期所著的经典小说《追忆似水年华》的第一段，以下是第一段的法文原文：

> Longtemps, je me suis couché de bonne heure. Parfois, peine ma bougie éteinte, mes yeux se fermaient si vite que je n'avais pas le temps de me dire: ‘Je m'endors.’ Et, une demi-heure après, la pensée qu'il était temps de chercher le sommeil m'éveillait; je voulais poser le volume que je croyais avoir encore dans les mains et souffler ma lumière; je n'avais pas cessé en dormant de faire des réflexions sur ce que je venais de lire, mais ces réflexions avaient pris un tour un peu particulier; il me semblait que j'étais moi- même ce dont parlait l'ouvrage: une église, un quatuor, la rivalité de François Ier et de Charles Quint.

很难承认，尽管努力学了10年，我对法语的理解还是十分有限，只能辨认出上文里一些奇怪的孤立的句子，如果没人帮我翻译，我根本看不懂这段文字。

以下是由专业翻译将它翻译成英文的结果[80]：

> For a long time I used to go to bed early. Sometimes, when I had put out my candle, my eyes would close so quickly that I had not even time to say ‘I'm going to sleep.’ And half an hour later the thought that it was time to go to sleep would awaken me; I would try to put away the book which, I imagined, was still in my hands, and to blow out the light; I had been thinking all the time, while I was asleep, of what I had just been reading, but my thoughts had run into a channel of their own, until I myself seemed actually to have become the subject of my book: a church, a quartet, the rivalry between François I and Charles V.[13]

这下好多了!但有趣的是，虽然这是一段优雅的英文，但它并不是那么直白好懂，至少对我来说是这样。当作者写下“I. . . seemed actually to have become the subject of my book: a church, a quartet, the rivalry between François I and Charles V”(直译：我……似乎真正成了这本书的主角：一个教堂、一出四重奏、弗朗索瓦一世和查理五世的竞争)时，到底指的什么？你怎么能“become”(成为)一个“church”(教堂)？他说的“quartet”(四重奏)又是什么意思？还有François I(弗朗索瓦一世)和Charles V(查理五世)之间有什么“rivalry”(竞争)？另外，对一个使用电灯的人而言，

“blow out the light”(吹灭烛火)又是什么意思？

接下来我们看看谷歌是怎么翻译这一段的：

> Long time, I went to bed early. Sometimes, when my candle, my eyes would close so quickly that I had no time to say: ‘I fall asleep.’ And half an hour later the thought that it was time to go to sleep would awaken me; I wanted to ask the volume that I thought I had in my hands and blow my light; I had not ceased while sleeping to reflections on what I had read, but these reflections had taken a rather peculiar turn; I felt that I myself was what spoke the book: a church, a quartet, the rivalry between Francis I and Charles V.[14]

谷歌翻译做的是一件很复杂的事情，跟专业的人工翻译工作类似。但你并不需要相当专业的翻译知识或者文学素养，就能够看出这段翻译其实挺烂的。在英语中，“blow my light”(直译：吹出我的光芒)这个短语毫无意义，这就让后面的句子显得也没有任何意义。事实上，这些句子读起来特别滑稽。而且翻译结果中包含了母语为英语的人永远不可能使用的短语。我们得到的总体印象是：这段文本大致可以辨认出是什么意思，但是行文扭曲、不自然。

当然，我们给谷歌翻译出了一道难题——翻译普鲁斯特的小说对一个专业的法译英译者而言都是个巨大的挑战。现在问题来了，为什么自动翻译工具这么难以处理文本呢？

关键就在于，你仅仅是懂得法语并不代表就能做好普鲁斯特小说的翻译。哪怕你精通法语，但普鲁斯特的小说仍然会让你摸不着头脑，不仅仅因为他的文字风格，要正确翻译他的小说，你就得理解它，这就需要你有大量的背景知识。关于20世纪初期法国社会和法国人生活的知识(例如你得知道他们使用蜡烛照明)，法国历史的知识(例如你得知道弗朗索瓦一世和查理五世之间的斗争史)，20世纪早期法国文学常识(例如当时的写作风格，还有作者可能引用的典故)，以及对普鲁斯特本人的了解(例如他最想表达的是什么)。谷歌翻译所使用的神经网络里可没有这些知识。

要理解普鲁斯特的小说需要各种各样的相关知识，察觉到这一点并不新鲜。我们在第三章提到的Cyc项目中就遇见过。还记得Cyc项目的目标是创建“包罗万象的知识库”， Cyc的假设是，这将是创造通用人工智能的基础。基于知识的人工智能研究人员肯定希望我向你们指出，早在几十年前他们就预见到这个问题了(来自神经网络研究界的尖锐反驳就是：基于知识的人工智能界根本没创造出来适用解决这个难题的技术，

对不对？)。但是，仅仅改进深度学习的技术就能解决这个问题吗？我认为并不是这样。深度学习将解决问题方案的一部分，我认为，一个合理的解决方案需要的不仅仅是更庞大的神经网络、更强大的处理能力，或者更多无聊的以法国小说形式出现的训练数据。它需要突破现有的模式，需要至少和深度学习本身一样闪亮的突破性进展。我怀疑这些将需要明确的知识表述方式，也需要深度学习：我们必须消除明确表示知识的世界和深度学习以及神经网络的世界之间的隔阂。

重大分裂

2010年，我应邀组织一个大型国际人工智能会议——欧洲人工智能大会(ECAI)，大会在葡萄牙里斯本举行。参加类似ECAI这样的会议是人工智能研究人员生活的重要部分，我们把自己的研究成果写下来，提交给大会，由大会的项目委员会审核。项目委员会通常是由相关领域著名科学家组成的小组，他们决定哪些研究成果值得在大会上发表。权威的会议大概只会接受五分之一的投稿，所以研究成果被大会接受是一件非常有意义的事情。真正大型的人工智能会议能够吸引超过5000份的投稿。所以，你能想象，被邀请担任ECAI的主席，我感到非常荣幸——这意味着科学界信任你，另外跟学院提升职加薪的时候也是值得大书特书的一笔。

作为主席，我的工作包括召集项目委员会成员，我非常希望能有来自机器学习研究领域的代表。但意外的事情发生了：每一位我试图邀请加入项目委员会的机器学习领域专家，都礼貌地拒绝了我。被婉拒是常见的事情——毕竟，这是一项艰巨的工作。但是我连一个人都找不到就很奇怪了，是我的问题吗？还是ECAI的问题？或者别的问题？

我向以前组织过这项活动的同事和组织过其他类似会议的人请教，他们也提到了同样的情况。机器学习研究领域似乎对所谓的“主流人工智能”事件不感兴趣。我知道机器学习领域的两件学界大事是神经信息处理系统(NeurIPS)会议和国际机器学习会议(ICML)。人工智能多数分支领域都有自己的专家会议，所以该领域的专家们更注重这类会议，这不足为奇。但在此之前，我根本没意识到，机器学习研究领域的许多人根本就不把自己视为“人工智能”的一部分。

事后来看，人工智能和机器学习之间的分裂似乎很早就有端倪了，也许是从1969年明斯基和帕普特出版了《感知器》这本书开始。正如我

们之前所提到的，这本书似乎在扼杀神经网络人工智能研究方面起到了重大作用，从60年代末一直到80年代中期PDP(并行分布模型)的出现。即使在半个世纪之后的今天来回顾，人们对这本书出版的后果仍然感到痛心。不管分裂的起源是什么，事实就是，在某种程度上，机器学习研究领域的许多人脱离了主流人工智能，沿着自己的轨迹发展。当然，也有许多研究人员认为自己可以轻松跨越机器学习和人工智能之间的隔阂。但直至如今，如果你给不少机器学习专家的研究工作贴上“人工智能”标签，他们会感到惊讶，甚至恼火：因为对他们来说，人工智能只是我在本书其他地方记录的一长串失败的想法罢了。

[1] 传闻汉诺塔来自印度的古老传说，印度教的主神梵天在创造世界之时，在世界中心贝拿勒斯(在印度北部)的圣庙里设置了三根柱子和64个金环，并设定了移动规则。当僧侣按照规则把所有的金环都移动完成时，世界将在一声霹雳中毁灭，梵塔、庙宇和众生都将同归于尽。

[2] P代表多项式时间，在计算机术语中，如果一个问题能够在多项式时间内解决，这个问题就是有意义可解的，即P问题，简单地说就是P问题的有效解决方案不会引起组合爆炸。NP问题指的是不确定能否在多项式时间内解决，但是确定能够在多项式时间内验证某个解是否有效的问题。是否能证明NP问题都可以等同于P问题，是当今计算机科学面临的一大难题，即P与NP问题。

[3] 人们将对于计算机来说最困难的问题，非正式地称为“AI完全”(AI-complete)或者“AI困难”(AI-hard)，以此说明解决了这些计算性问题就相当于解决了人工智能的核心问题——让计算机和人类或者强人工智能一样聪明。将一个问题称为“AI完全问题”，意味着它不能被一个简单的特定算法解决。

[4] 顺势疗法是替代医学的一种，其理论基础是“同样的制剂治疗同类疾病”，意思是为了治疗某种疾病，需要使用一种能够在健康人中产生相同症状的药剂。

[5] MYCIN系统，是一种帮助医生对住院的血液感染患者进行诊断和选药治疗的人工智能。

[6] 1英寸≈2.54厘米。

[7] 在层次系统中，上一层次单元所具有的构成其下一层次单元所不具有的某些性质。这种性质往往是由于下一层次单元及其相互联结方式的非线性性质产生，也即总体不等于其各个部分之和。

[8] 布鲁克斯是iRobot公司的创始人，该公司广受欢迎的产品Roomba智能扫地机器人就来自他的研究成果[81]。

[9] 博弈论的英文为game theory，直译为“游戏理论”。

[10] 信用分配问题(credit assignment problem)，也有译为赞誉分布、功劳分配

的，通俗来讲可以比喻成你吃了10个包子后吃饱了，但是你并不知道具体是哪个包子为你吃饱的贡献比较大。

[11] 异或是一种逻辑运算，计算机符号为“XOR”，运算法则为当a、b两个值不相同时，异或结果为1，当a、b值相同时，异或结果为0。

[12] 鳕鱼系统是世界领先的国际象棋程序之一。

[13] 中文翻译如下：我每天都早早躺下，已经持续很长一段时间。有时候，蜡烛刚灭，我甚至来不及咕哝一句“我要睡着了”，就进入梦乡。半小时之后，我才想起应该睡觉，这一想反倒让我清醒过来。我准备把感觉还握在手里的书放好，吹灭灯火。一直到睡着，我都在思考刚才读的那本书，只是思路有点特别：我总觉得书里说的事儿，什么教堂呀，四重奏呀，弗朗索瓦一世和查理五世争强斗胜呀，全都同我直接相关。

[14] 法文原文使用翻译软件翻译结果如下：长久以来，我睡得很早。有时候，我的蜡烛刚刚熄灭，我的眼睛闭得太快，以至于我没有时间对自己说：“我睡着了。”半小时之后，想到是寻找睡眠的时候了，我醒来了。我想放下我还以为我手上还有的书卷，吹出我的光芒。我在睡觉的时候一直在思考我刚才读到的东西，但这些思考却有一点特别，我觉得我就是作品中提到的那个人：一个教堂，一个四重奏，弗朗索瓦一世和查理五世的竞争。

第三部分 我们将去向何处

第六章 人工智能的今天

深度学习为人工智能的应用打开了大门，在21世纪的第二个10年里，自从20世纪90年代万维网出现以来，还没有另一门新技术如人工智能这样吸引人们的注意力。每个拥有数据和需要解决问题的人都忍不住要问，深度学习是否能够帮助他们——在很多情况下，答案是肯定的。人工智能已经出现在我们生活的方方面面，我们随时都能感受到它的存在。但凡涉及技术的领域，都有人工智能的身影：教育、科学、工业、商业、农业、医疗保健、娱乐、媒体和艺术等各个领域。未来，或许有些领域会有非常明显的人工智能痕迹，有些领域则不会。人工智能将悄然无声地嵌入我们的世界，就像今天的计算机一样。正如计算机和万维网，人工智能也将改变我们的世界。我无法为你列举出人工智能的全部应用，就像我无法列举所有的电脑软件一样。但我会跟你分享一些我喜欢的过去几年中出现的案例。

你可能还记得，2019年4月，世界上第一张黑洞的照片问世[82]。在一次令人难以置信的实验中，天文学家利用分布在世界各地的八台射电望远镜收集的数据，构建出一个直径400亿英里[1]、距离地球5500万光年的黑洞图像。这幅图像是本世纪最引人注目的科学成就之一。但你可能不知道的是，黑洞图像是通过人工智能实现的：人们使用先进的计算机视觉算法来重建图像，“预测”图片中缺失的元素。

2018年，来自计算机视觉处理器公司英伟达的研究人员证明了人工智能软件能够创造出虚假的人物照片，并且能够完全令人相信它是真实的[83]。这些图片由被称为生成性对抗网络的全新神经网络制造。照片看起来简直是不可思议：乍一看，它们十分逼真，很难相信那不是真人。我们的眼睛告诉我们，这就是真实的照片——但事实上它们是由神经网络创造的。这种技术在21世纪刚开始的时候难以想象，而在未来它会成为虚拟现实(Virtual reality， VR)的关键组成部分：人工智能正在构建令人分辨不清的虚拟现实。

2018年末，深度思维的研究人员在墨西哥的一次会议上公开发布了

AlphaFold，这是一个能根据基因序列来预测蛋白质3D结构的程序，能够解决名为“蛋白质折叠”的问题[2] [84]。蛋白质折叠问题包括预测蛋白质分子的形状，了解它们将对治疗类似阿尔茨海默病等疾病有极其重要的意义。不幸的是，这个问题非常艰难。而AlphaFold使用经典的机器学习技术来学习如何预测蛋白质结构，这意味着在理解蛋白质性状方面迈出了一大步。

在这一部分，我想详细讲述人工智能应用最突出的两个方面：第一个是人工智能在医疗领域的应用，第二个是长久以来无数人的梦想——无人驾驶汽车。

人工智能医疗

> 人们应该停止培养放射科医生了。很明显，在五年之内，深度学习的系统会比放射科医生做得更好。
>
> ——杰夫·辛顿(2016)

> Cardiogram可以担任您的个人健康助理。我们会把可穿戴设备变成一个持续的健康监测仪，不仅可以用来追踪睡眠和健康状况，而且或许在未来的某一天，可以预防中风、挽救您的生命。
>
> ——Cardiogram公司网页[85]

哪怕是对政治和经济最不敏感的人都能意识到，不管是对政府还是对企业而言，医疗卫生都是全球最重要的经济问题之一。一方面，在过去的两个世纪里，医疗卫生的改善可能是工业化和科技世界带给人们的最大福祉。1800年，欧洲人的寿命预测不会超过50岁[86]，而今天，预测一个人能活到70岁是非常合理的；在发达国家，产妇因分娩而死亡的情况十分少见。这些巨变一定程度上是因为人们对卫生有了更好的认识，但不可否认，药物研究和疾病治疗研究的飞速发展也起到了同样重要的作用。尤其是20世纪40年代抗生素的出现，首次为细菌感染提供了可靠、有效的治疗方法。当然，这些进步目前还没有覆盖全球所有地区。在我撰写本文的时候，中非共和国的国民预期寿命仍然只有51岁，世界上还有许多地方，分娩对母亲和孩子而言都不啻闯一次鬼门关。但是，总的来说，发展趋势是积极的，这当然值得庆贺。

但这些可喜的进展又带来了新的问题。首先，人类的平均寿命增长了，老年人通常比年轻人需要更多的医疗卫生资源，这就意味着医疗卫生的成本在整体上升。第二，随着我们开发更多的治疗疾病的新方法和

研制更多新的药品，可以治疗疾病的总体范围增加了，这也导致了更多的医疗支出。当然，医疗费用昂贵的一个根本原因是，提供医疗保障服务所需要的资源昂贵，具备相应技能和资质的人也很少：在英国，要成为一名合格的全科医生，至少要培训10年。

由于这些问题，医疗卫生——尤其是相关的资金——向来都是政客们争论不休的长期问题。在英国，国家医疗服务(NHS)于20世纪40年代作为一项国家服务建立，通过税收支付，目的是为每个人提供免费的医疗保障。而关于如何资助国家医疗服务的争论从来就没停止过。我们都喜欢国家医疗服务，但都无休止地争论应该给它多大程度的资助。

医疗保障至关重要，但实现起来也困难重重。那么，如果有一项技术能够解决问题，那简直就太美妙了，不是吗？

人工智能应用在医学领域也不是新鲜事了，我们在之前就了解过MYCIN专家系统，它在诊断人类血液疾病病因方面表现得比人类医生更专业，因此广受赞誉。早在20世纪80年代初，医疗卫生资金就跟现在一样成为令人头疼的难题。出于之前讲述的原因，制造出能够捕捉医疗专业能力的专家系统程序，这个想法让人兴奋不已。所以在MYCIN以后，类似的医疗卫生系统如雨后春笋般涌现出来，也没什么好奇怪的。不过公平地说，几乎没有多少系统在离开实验室以后还能起作用。不过这些年，人们对人工智能用于医疗的兴趣开始报复性反弹，有新的进展表明，人工智能在医疗应用方面前途无量。

个人医疗健康管理系统是人工智能在医疗健康领域的重要新机遇。可穿戴设备——以Apple Watch为代表的智能手表，还有以Fitbit为代表的运动健身智能手环之类的出现，使个人医疗健康管理成为可能。这些设备持续监测我们的生理数据，比如心率和体温等。大量的用户不断生成当前健康状况的数据流，人工智能系统可以对这些数据进行分析，不管是本地的系统(比如你的智能手机)还是将数据上传至互联网上的人工智能系统。

千万不要低估这项技术的潜力，这是有史以来第一次，我们能够对自己的健康状况进行持续监测。在最基本的层面上，基于人工智能的医疗健康管理系统能够为我们的健康管理提供合理的建议。从某种意义上来说，这正是智能手环、智能手表正在做的事情：它们监测我们的运动，还可以为我们设定目标(“每天万步行走挑战赛”就是一个明显的成功

案例)。经验证明，我们可以通过将运动目标进行游戏化的方式来提高人们对目标的遵从性。将目标转化为竞赛或者游戏，还可以借助社交媒体来进行交互。

大众市场的可穿戴设备还处于初级阶段，但有许多迹象表明未来的发展潜力。2018年9月，苹果推出了第四代Apple Watch，首次包含了心脏监护仪。手表上的心电图应用程序可以监控心率跟踪器提供的数据，并且能够识别心脏病的症状。如果有必要，甚至可以为使用者呼叫救护车。应用程序可以监测心房颤动难以捉摸的迹象——不规则的心跳，这可能是中风或其他突发性疾病的征兆。智能手机中的加速计应用可以用来识别坠落，如果需要，可以代为呼叫救援。这样的系统只需要相当简单的人工智能技术：现在我们可以随身携带一台功能强大的智能手机，它可以保持互联网连接，并且可以连接到配备了一系列生理传感器的可穿戴设备上。

某些个人健康管理应用甚至不需要传感器，只需要一台智能手机。我在牛津大学的同事们相信，仅仅从使用智能手机的方式就可以检测出阿尔茨海默病的发病征兆。人们使用手机方式的改变，或者手机记录的行为模式发生改变，或许都是某些疾病的前兆。这些征兆大多发生在人们注意到病人有明显症状、医生正式下达诊断之前。阿尔茨海默病是一种毁灭性的疾病，对人口老龄化的社会构成了巨大挑战，可以辅助早期诊断或者管理阿尔茨海默病的工具将非常受欢迎。当然，这项工作还处于起步阶段，但它至少提供了未来的一种可能性。

这些新技术的出现令人兴奋，它们带来了机遇，同时也带来了潜在的隐患，其中最明显的就是隐私问题。可穿戴设备跟人进行亲密接触，它不断监视我们，虽然它获取的数据能够给我们提供帮助，但也为数据滥用埋下了隐患。

保险业是一个值得关注的领域。2016年，健康保险公司Vitality开始随保险单附送苹果手表。手表监视你的运动情况，然后根据你的运动状况设定保险费用承诺。如果，某个月，你决定不做任何运动，就躺在沙发上当个懒虫，你可能需要支付全额保费，但你也可以在下个月通过疯狂节食来抵扣，以降低保费。这样的计划或许没什么直接的问题，但它确实暗示了一些令人不安的情况。例如，2018年9月，美国约翰汉考克人寿保险公司宣布，今后只向准备佩戴追踪运动数据的可穿戴设备的个人提供保险单[87]。这一声明引来了大众的批评。

进一步说，如果我们想获得国家医疗保障计划(或者其他国家福利)，就必须接受监督并且达到锻炼目标，这又将如何？你想要医疗保险？那每天走一万步再说!有些人会觉得这没什么问题，但对另一些人来说，这是对我们基本人权的严重侵犯和对监测数据的滥用。

自动化诊断是人工智能在医疗保健领域的又一个令人兴奋的潜在应用。在过去的10年里，诸如X射线机和超声波扫描仪之类的医疗成像设备，其成像数据用机器学习的方式来进行分析，受到了极大关注。在撰写本书的时候，有一篇新的研究成果发表，文章表示人工智能系统可以有效识别医学图像中的异常。这是机器学习的一个经典应用：我们通过正常的图像和异常的图像示例来训练机器学习程序，最终的目的是让程序能够识别出图像中的异常。

在这项领域中有一个广为人知的案例。2018年，深度思维公司宣布正在与伦敦的摩菲眼科医院合作，开发从眼部扫描检查中自动识别疾病和异常的技术[88]。眼部扫描检查是摩菲眼科医院的主要检查之一，他们通常每天要做1000次眼部扫描检查，分析扫描检查结果是医院工作中的一个重要部分。

深度思维的系统使用了两个神经网络，第一个用于“分割”扫描图像(识别图像的各个部分)，第二个用于诊断。第一个网络大约训练了900个图像，学习人类专家如何对扫描图像进行区分识别。第二个网络训练了大约15 000个案例。实验表明，该系统的性能已经达到甚至超过了专家水平。

这些结果都很好，你也可以随手在网上搜出一大堆其他的引人注目的例子，说明当前的人工智能技术是如何被用来建立具有类似能力的系统的——在X光片上识别恶性肿瘤、通过超声扫描诊断心脏病等等。杰夫·辛顿，你可能还记得他是非常成功的图像识别程序AlexNet的创立者之一，他非常确信机器学习将为医学影像诊断提供解决方案，因此他对放射科医生做了一个相当大胆的声明——就在这一小节开始的时候，我引用过。不出所料，这激怒了放射科的医生，很快就有人指出，当好一个放射科医生所需要的技能可不仅仅是看X光片[89]。

也有不少人认为，我们需要谨慎地推动人工智能在医疗领域的应用。首先，医疗卫生行业是一个人文学科，可能比起任何职业都更需要与人互动和与人交往的能力。全科医生需要“解读”病人，了解病人的社

会背景，了解哪些治疗方案对这个病人可能更有效，而哪些方案是无效的，等等。所有证据表明，我们目前建立的人工智能系统，在分析医疗数据方面确实已经达到人类专家的水准，但这只是人类医疗工作者工作的一小部分(尽管是非常重要的一部分)。

另一个反对在医疗卫生行业广泛应用人工智能的论点认为，有些人更倾向于依赖人的判断，而不是机器的判断。他们愿意和人打交道，这里有两个问题需要说明。

首先，把人类专家的判断奉若圭臬，实在是太过天真的想法。每个人都会有缺陷，即使是最勤奋、最有经验的医生，也会有感到疲惫或情绪化的时候。而且，不管我们怎么努力去克服，都难免或多或少带有偏见以及经验主义。另外，我们人类并不太擅长做理性决策，而机器可以做出与人类专家同等水准的判断，医疗卫生行业的挑战或者说机遇，应该是将机器的这种能力用最佳的方式利用起来。我的信念是，人工智能的作用并不是取代人类的医疗卫生专业人员，而是用来增强他们的能力，让他们从某些烦琐的工作中解脱出来，更专注于专业领域中真正困难的部分；以及，提供另一种角度的观点以供参考，让他们的思考更加全面。

其次，在我看来，选择人类医生还是人工智能医疗程序那是发达国家的人才会抱怨的问题，对世界上许多地方的人来说，要么接受人工智能，要么就无人处理。在专业医生极度缺乏的地区，人工智能可以做很多事情。它让获取医疗卫生资源困难地区的人们有了更多的希望，这才是真正最令人激动的前景。在人工智能给予我们的所有机遇中，这可能是会产生最大社会影响的一个。

无人驾驶汽车

> 制造比空气还重的飞行机器是不可能成功的。
>
> ——开尔文(Kelvin)勋爵，皇家学会会长，1895年

在撰写本文的时候，全球每年有超过100万人死于和汽车有关的交通事故，仅中国和印度就占了其中的四分之一；每年还有5000万人在跟汽车有关的交通事故中受伤。这些数字触目惊心，想象一下，如果出现一种每年可以夺取100万人性命的新型流感病毒，那一定会引起全球性恐慌。然而，我们却习惯了公路上的危险——我们似乎已经接受这就是

现代社会的现状。不过，人工智能能够带来大幅降低交通事故的前景：在发展智能中期内出现无人驾驶汽车，已经成为可能。最终，它能够拯救无数人的生命。

当然，无人驾驶汽车还有许多其他好处。利用计算机程序来控制汽车驾驶，显然更高效，能够更好地利用稀缺和昂贵的燃料或动力资源，从而产生更环保、运行成本更低的汽车。计算机程序在利用电子地图和导航方面也更具备优势，例如，可以为拥挤的交通道路带来更好的通行能力。如果汽车变得安全，它们就不再需要如此昂贵和沉重的保护底盘，这将再次降低汽车的价格和油耗。甚至有一种观点认为，无人驾驶汽车会减少私家车的拥有量，因为无人驾驶的出租车又便宜又方便，那么，拥有自己的私家车在经济上就没多大意义了。

由于上述以及更多原因，无人驾驶汽车显然是一个非常有前景的研发领域，因此，这一领域有着悠久的研发历史，也不足为奇。汽车在20世纪20至30年代成为大众市场的产品，由它引发的伤亡规模——主要是驾驶员人为操作失误——立刻引发了人们对汽车自动驾驶可能性的讨论。虽然自20世纪40年代以来，人们在自动驾驶领域就不断地尝试，但直到70年代微处理器技术出现以后，它们才真正变得可行。不过，无人驾驶汽车面临的挑战也是艰巨的，最根本的问题就是感知。如果你能够找到一种方法，让一辆汽车能够随时准确地知道它自己在哪里，周围环境是怎么样，那么恭喜你，你已经找到解决无人驾驶问题的方法了。而要解决感知问题，我们需要采用现代机器学习技术：没有它们，无人驾驶汽车无法实现。

由欧洲泛政府研究组织欧洲研究协调局(EUREKA)出资赞助的普罗米修斯工程，是无人驾驶汽车技术的先驱。普罗米修斯工程从1987年延续到1995年，并在1995年进行了一次示范表演。一辆汽车在无人驾驶的情况下从德国慕尼黑开到丹麦的欧登塞，然后返回。虽然平均5.5英里就需要进行一次人工干预，但是在没有人工干预的情况下，最长的一次无人驾驶距离约为100英里。这是一个了不起的壮举，鉴于当时的计算机有限的计算力，这个成就更加令人惊叹。虽然普罗米修斯工程只是为了证实无人驾驶的概念是可以成为现实的，它仍然引导了不少现代商用汽车创新的风向，比如智能巡航控制系统。最重要的是，普罗米修斯工程昭示了这项技术可以商业化。

2004年，美国国防高级研究计划局(DARPA)组织了一场无人驾驶汽

车顶级挑战赛，邀请研究人员组队参加挑战赛，让无人驾驶的车辆穿越150英里的美国乡村。共有106支来自大学研究院和汽车公司的参赛队，每个参赛队都渴望能赢取第一名的100万美金DARPA大奖。最后，有15支参赛队伍进入决赛，但在这一届比赛中，没有一支队伍完成超过8英里的赛程，有的车辆甚至都没能驶离出发区。跑得最远的是来自卡内基-梅隆大学的无人驾驶汽车，虽然它仅仅前行了7.5英里就偏离了航道，卡在堤坝上。

我对这件事情的记忆是这样的：大多数人工智能研究人员都把2004年的挑战赛作为证据，证明无人驾驶汽车技术离实际应用还有一段距离。听说计划局立即宣布在2005年的竞赛中奖金翻一番，达到200万美元，让我有点吃惊。

2005年的比赛吸引了更多的参赛者，总共有195支队伍参赛，最终闯入决赛的有23支。决赛在2005年10月8日举行，挑战目标是让无人驾驶车辆穿越132英里的内华达沙漠。这次，有5支参赛队伍完成了挑战。最终的冠军被斯坦福大学研究团队摘得，获得了200万美元的奖金。斯坦福大学获胜的汽车名叫**STANLEY**，由塞巴斯蒂安·特隆(Sebastian Thrun)带领的团队打造。它是由一辆大众途锐汽车改造的，用了不到7个小时就完成了比赛，平均时速20英里。STANLEY配备了7台车载电脑，它们需要解读GPS、激光测距仪、雷达和视频传输中的数据。

2005年的无人驾驶汽车挑战赛是人类历史上最伟大的科技成就之一。从那一天开始，无人驾驶汽车成了已经解决的问题，就像一个多世纪以前，比空气更重的飞机成为已经解决的问题那样。莱特兄弟的第一次飞行只持续了12秒，在这段时间里，飞行器离地面只有120英尺。但那12秒的飞行历程证明了，比空气更重的飞机已经成为现实——2005年的挑战赛以后，无人驾驶汽车也如此。

紧随其后的是一系列挑战赛，其中最重要的可能是2007年的无人驾驶汽车城市挑战赛。2005年的比赛是在乡村道路上进行的，而2007年则是在城市环境中。无人驾驶汽车必须完成整个赛程，同时遵守加利福尼亚州道路交通法规，并应对停车、十字路口通行和交通拥堵等日常情况。有36支队伍参加了全国预选赛，其中11支队伍闯入决赛。2007年11月3日，在南加州一个废弃机场，决赛开始了。6支队伍成功完成了挑战，获胜者来自卡内基-梅隆大学，在4小时内跑完全程，平均车速每小时14英里。

从那以后，无人驾驶汽车技术领域获得了大量投资，既有老牌汽车公司不顾一切地拒绝被时代车轮甩掉，也有新公司发现有机会来抢夺传统汽车制造商的蛋糕。

2014年，美国自动化工程师协会提供了一个实用的分级方案，为自动驾驶分级制定了标准[90]：

Level 0：人工驾驶。汽车没有自动控制功能，驾驶员始终自主控制车辆(尽管车辆会提供警告和其他数据用来辅助驾驶员驾驶)。如今公路上的绝大部分汽车都是L0级。

Level 1：辅助驾驶。汽车提供了一定程度的控制，通常是在常规驾驶方面，但驾驶员仍须在驾驶过程中保持全神贯注。自适应巡航控制系统是辅助驾驶的一个例子，它可以使用刹车和油门来控制车速。

Level 2：部分自动化。在这个等级，汽车可以参与转向和速度的自主控制，但驾驶员同样需要持续监控驾驶环境，并准备好在必要的时候进行干预。

Level 3：有条件自动驾驶。在这个等级，驾驶员已经不再需要持续监控驾驶环境，尽管汽车可能会要求用户在遇到无法应对的情况下进行控制。

Level 4：高度自动化驾驶。在这个等级，汽车能够自动完成正常驾驶操作，不过驾驶员仍然可以干预驾驶行为。

Level 5：全自动驾驶。你只需要坐上一辆车，说出你的目的地，然后剩下的所有事情都交给汽车自动处理。这种汽车甚至连方向盘都没有。

在撰写本书时，最先进的商用无人驾驶汽车系统可能是特斯拉的自动驾驶系统，最初出现在特斯拉S型车上。2012年发布的特斯拉S型是高规格电动车系列中的旗舰车型，在发布之时，它或许是全世界技术最先进的市售电动车。从2014年9月起，所有特斯拉S型车都配备了摄像头、雷达和声程传感器。2015年10月，特斯拉发布了新款车用软件，启用了这些高科技套件，完成“自动驾驶”功能——当然，是一种有限的自动驾驶能力。

媒体立刻开始盛赞自动驾驶仪实现了第一款无人驾驶汽车，尽管特斯拉不厌其烦地指出了这项技术的局限性。特斯拉特别强调，当自动驾驶仪接通时，驾驶员应始终把手放在方向盘上。就上述自动驾驶分级而言，特斯拉的自动驾驶仪似乎处于L2级。

无论这项技术有多先进，很明显，涉及自动驾驶仪的严重事故还是会发生，而第一例特斯拉自动驾驶仪导致人员死亡的事故很快成为全世界的头条新闻。2015年5月，佛罗里达州的一位特斯拉车主在路上与一辆18轮卡车相撞，不幸身亡。有报道称，汽车的传感器被白色卡车在明亮的天空下的景象给迷惑了，结果汽车的人工智能系统未能识别出路上

还有另一辆车存在，直接高速撞上卡车，司机当场死亡。

其他的事件也凸显了目前无人驾驶技术的一个关键问题。在L0级人工驾驶模式下，对驾驶员的期望是非常清晰的：驾驶员必须掌控车的一切。而在L5级的全自动驾驶模式下，同样很清晰：驾驶员什么也不用做。但是，在这两个极端之间，对驾驶员的期望就相对比较模糊了。佛罗里达州的事故和其他类似事故表明，驾驶员似乎对自动驾驶技术期望过高：把它当成L4或者L5级的自动化驾驶系统，事实上它的功能远远没有达到这个等级。驾驶员期望的功能和自动驾驶仪实际能够实现的功能之间如此不匹配，至少在一定程度上是媒体过于兴奋地鼓吹所造成的，它们似乎不太能理解和传达技术等级能力的微妙之处(特斯拉给这个系统命名为“自动驾驶仪”也得负点儿责任)。

2018年3月，另一起涉及无人驾驶汽车的事故引发了人们对这项技术的进一步质疑。2018年3月18日，在亚利桑那州的坦佩市，优步公司的一辆无人驾驶汽车在无人驾驶模式下撞死了49岁的行人伊莱恩·赫茨伯格(Elaine Herzberg)。与这类事故的典型情况一样，引发事故的原因是复杂的。这辆车当时的速度已经超过了自动紧急制动系统能够处理的速度，所以当汽车意识到需要紧急制动的时候，已经来不及了。虽然汽车传感器意识到前方有“障碍物”(受害者伊莱恩·赫茨伯格)，需要紧急制动，但软件似乎规避了这个行为(在程序员看来，这似乎在说明软件系统有缺陷，至少是存在优先级处理不当的问题)。而且，最重要的是，车上的“人类驾驶员”应该干预这样的事情，可她似乎一直在用智能手机看电视，对外界环境关注甚少。很可能她盲目信任了汽车的无人驾驶能力。伊莱恩·赫茨伯格不幸遇难的悲剧是完全可以避免的：这个问题是人为的，而不是技术上的。

令人沮丧的是，在看到实用的、大众化的无人驾驶汽车之前，还会有更多这类悲剧发生。我们需要尽己所能合理地预测和避免这类悲剧，但无论如何，它们总是会发生的。当它们发生以后，我们需要从中吸取教训。就像飞机的飞行控制器发展史告诉我们的，从长远来看，总会出现更加安全的交通工具。

目前围绕无人驾驶汽车的一系列活动表明，这项技术已经日趋成熟，但是离它进入我们的现实生活还有多久呢？我们什么时候可以跳上一辆无人驾驶汽车，只需要说出自己的目的地，就可以轻松抵达呢？在这方面做得最公正和权威的是美国加利福尼亚州的无人驾驶监管机构。

无人驾驶汽车公司必须向该机构提供详细的相关信息，才能获取在加州公共道路进行无人驾驶汽车测试的许可证，其中最重要的信息是自动驾驶汽车脱离报告。脱离报告的内容说明公司相关车辆在无人驾驶情况下行驶的英里数，以及在测试期间发生过多少次脱离接触。脱离接触是指人类驾驶员不得不干预汽车行驶，接管汽车控制权的情况——这是伊莱恩·赫茨伯格的悲剧中驾驶员应该做的事情。脱离接触并不意味着如果没有人工干涉就一定会出现事故(更别提死亡事故了)，但它仍然是衡量自动驾驶技术性能的标准。自动驾驶时，每千英里中出现脱离接触的次数越少越好。

2017年，有20多家公司向加利福尼亚州提交了自动驾驶汽车脱离报告。从行驶里程数和每千英里最低脱离次数来看，一家名为Waymo的公司遥遥领先，该公司的自动驾驶汽车平均每行驶5000英里才报告1次脱离。表现最差的是汽车巨头梅赛德斯-奔驰，每千英里不少于774次脱离。Waymo是谷歌旗下的无人驾驶汽车公司，最初，它是谷歌内部的一个项目，运营负责人是塞巴斯蒂安·特隆，他曾经率领团队赢得2005年美国国防高级研究计划局组织的无人驾驶汽车挑战赛。2016年，Waymo成为谷歌的子公司，2018年，Waymo的报告说，该公司旗下无人驾驶汽车已经达到平均行驶超过11 000英里才报告1次脱离的水准。

那么，这些数据告诉我们什么呢？尤其是，无人驾驶汽车还要多久才能进入我们的日常生活？

好吧，从宝马、梅赛德斯-奔驰和大众等传统汽车巨头相对较差的表现中，我们可以得出的第一个结论是：汽车行业经验的积累并不是无人驾驶汽车技术取得成功的关键条件。仔细想想这并不奇怪：无人驾驶汽车的关键不是内燃机，而是软件——人工智能软件。因此，美国汽车巨头通用公司在2016年收购了无人驾驶汽车公司Cruise Automation，金额保密(但显然数额巨大)，而福特公司给自动驾驶初创公司Argo AI投资10亿美元。两家公司都公开了推出无人驾驶汽车的雄心勃勃的声明：福特公司预测将在2021年前投入运营一款“完全自动驾驶”的商用汽车[91]。

当然，我们并不知道汽车公司各自采用的在何时需要进行脱离接触的确切标准，也许梅赛德斯-奔驰只是过于谨慎而已，但看上去我们不得不承认一个结论：至少在撰写本书的时候，Waymo公司遥遥领先。

我们将脱离接触和人类驾驶员安全行驶做个有趣的对比。对于后者，虽然没有明确的统计数据，不过在美国，人类驾驶员出现严重事故的平均驾驶里程约为10万英里，甚至100万英里。这就表明，即使是市场领导者Waymo，也必须将其技术提高两个数量级，才能达到与人类驾驶员相当的道路安全驾驶水准。当然，并不是Waymo报告的所有脱离都会导致事故，因此这种比较也不够科学，但至少可以说明，无人驾驶汽车公司现在仍然面临艰巨的挑战。

有趣的是，与无人驾驶汽车技术工程师交谈会发现，他们认为这项技术的关键难点在于如何应对突发事件。我们可以训练汽车应对大多数可能出现的危险，但当汽车遇见一种与训练中任何事件都不同的情况，会发生什么呢？虽然大多数驾驶场景都是常规和可预期的，但难免会出现完全无法预计的突发状况。在这样的情况下，人类驾驶员可以凭借丰富的经验进行处理，利用经验思考处理方案，如果实在来不及思考，也会凭借直觉处理。而无人驾驶汽车没有直觉这种奢侈的东西——在可以预见的未来，它们也不会拥有。

另一个艰难的挑战是如何从我们的道路现状(道路上所有车辆都是由人类驾驶)，到一个混合过渡(即道路上的车辆部分由人类驾驶，部分自动驾驶)，最终过渡到完全无人驾驶。一方面，无人驾驶汽车在驾驶时的行为与人类不同，这会让与之共用道路的人类驾驶员感到困惑和不安。另一方面，人类驾驶员的行为是不可预测的，他们不一定会严格遵守交通法规，这就使人工智能很难理解他们的行为，并与之安全地互动。

鉴于我对无人驾驶汽车技术进步的乐观评价，以上的问题听起来可能让人十分悲观。所以，让我尽可能解释一下我认为未来几十年内，事情会怎样发展。

首先，我的的确确相信无人驾驶汽车技术在某种形式上很快就会在日常生活中应用——当然，是在未来10年内。然而，这并不意味着L5级的自动驾驶可以很快实现。相反，我认为，我们将开始看到无人驾驶技术在特定的“安全”领域开始推广，并逐渐走向更广阔的世界。

那么，这些技术将率先在哪些领域开始推广呢？我认为采矿业是一个很好的例子，也许从澳大利亚西部或加拿大阿尔伯塔省的大型露天矿开始：那里地广人稀，行人和喜欢骑自行车蛇行或做出其他危险动作的人也少得多。事实上，采矿业已经大规模使用自主汽车驾驶技术。例

如，横跨英国和澳大利亚的跨国矿业集团力拓集团于2018年宣称，在西澳大利亚皮尔巴拉地区，他们的大型自动卡车车队已经运送超过10亿吨矿石和矿产[92]。从公开的信息看，这些自动卡车离L5级别的全自动驾驶还差得远——更偏向“自动化”而不是“自主驾驶”。不过这是个很好的例子，说明无人驾驶汽车在有限的环境中能够发挥巨大作用。

同样，无人驾驶车辆似乎非常适用于工厂、港口或者军事设施区域。我相信在未来几年内，无人驾驶技术将在这些领域得到大规模应用。

除了这些特殊领域的应用，对于日常使用的无人驾驶技术，还有几种可能的情况，其中有一部分已经成为现实。或许在一些受限制的城市环境，抑或特定的路线上可以看到低速行驶的“微型出租车”。事实上，在撰写本文的时候，已经有几家公司在试用类似服务了，尽管使用范围受到严格限制(而且，至少目前，微型出租车车厢里面有人类“安全驾驶员”保驾护航，用以处理紧急情况)。在伦敦这样的城市里，这类低速行驶的车辆完全不是问题，因为众所周知，伦敦的交通状况总是很糟糕，行车速度一直都很缓慢。

另外一种可能性是在城市和主要高速公路上设置无人驾驶汽车专用车道。大多数城市已经设置有机动车道和自行车专用道，那为什么不设置无人驾驶汽车专用道呢？这样的车道可以通过传感器和其他技术来辅助自动驾驶车辆，无人驾驶汽车专用道的存在也会向与之共用道路的人类驾驶员发出一个明确的信号：小心机器人司机!

至于L5级别的全自动驾驶，恐怕离我们还有一段距离。但它终究会出现，我最乐观的预测是，从撰写本书之时往后20年，或许L5级别的全自动驾驶才能普及。我敢打赌，我的孙辈将觉得他们的祖父竟然要亲自开车上路，真是既可怕，又充满了乐趣。

第七章 杞人忧天——我们想象中的人工智能会出什么错

进入21世纪以后，人工智能的飞速发展引起了媒体的广泛关注。其中有一些报道是公正合理的，不过坦白地说，大部分报道都愚蠢得无可救药。其中有一些报道颇有知识性和引导性，而大部分则是杞人忧天式的恐吓。比如2017年7月，各大媒体争相报道脸书关闭了两个人工智能系统，因为它们开始用自己的人工智能语言进行交流(显然它们的设计者无法理解)[93]。当时的新闻头条和社交媒体报道立场明确地暗示，脸书关闭这两个系统是因为害怕它们失去控制。事实上，脸书的人工智能实验是常规化并且完全无害的，连人工智能专业的学生都能承担这类实验项目。脸书的系统策划疯狂杀人的可能性和你家的微波炉突然变身成杀手机器人的概率一样大，实话实说，那是根本不可能的。

一方面，我觉得对于脸书事件的报道相当滑稽，但另一方面，这也让我十分沮丧。问题在于，这样滑稽的报道迎合了大众对人工智能的“终结者式恐惧”：我们正在创造一些自己无法控制的东西，这会给人类的生存带来风险〔此处你应该能听到阿诺德·施瓦辛格(Arnold Schwarznegger)在《终结者》影片中经典的角色配音〕。当然，我们创造出怪物的想法绝不是现代才有的：它至少可以追溯到玛丽·雪莱的《弗兰肯斯坦》。

这种说法仍然主导着有关人工智能未来的争论，现在的人们还经常用讨论核武器的口吻来讨论人工智能的未来。亿万富翁企业家、贝宝公司(PayPal)和特斯拉公司的联合创始人埃隆·马斯克(Elon Musk)就对此担忧不已。他发表了一系列公开声明，表达了自己的担忧，并捐赠了1000万美金作为研究经费，支持负责任的人工智能研究。2014年，当今最著名的科学家斯蒂芬·霍金曾公开表示，他担心人工智能会成为人类生存的威胁。

鼓吹人工智能“终结者式恐惧”真的会带来很严重的后果，原因如下：首先，它让我们担心一些我们完全没必要担心的问题；其次，它让人们把注意力从真正应该关注的人工智能问题上转移开。这些问题可能不像终结者幻想那样吸引人的眼球，可以成为头条新闻，但它们才是我

们现阶段确确实实应该关注的。因此，在这一章中，我想解决有关人工智能的“终结者式恐惧”：类似《终结者》中的毁灭场景，到底有多大可能性出现；并且，人工智能是怎么出错的。接下来我会从直面这个故事开始，讨论它是怎么发生的，以及发生的概率有多大。这就会引出对人工智能伦理的讨论——人工智能系统充当道德智能体的可能性，以及已经提出的各种有关人工智能的伦理框架。最后，我将提醒大家注意人工智能的一个特点，就是它们很容易出故障，尽管还没到特别可怕的程度：如果我们想要一个人工智能系统代替我们工作，那么就需要跟它沟通我们想要的东西。但事实证明这很难做到，若是我们在传达意愿的时候稍有偏差，或许人工智能系统会给出我们所要求的、但并非我们真正想要的东西。

奇点主义？纯属胡扯!

在当代人工智能中，终结者式的场景通常跟名叫奇点的想法联系在一起，这个想法来源于美国未来学家雷·库兹韦尔(Ray Kurzweil)于2005年出版的著作《奇点临近》[94]。

> 奇点临近，其背后的关键思想是，人类创造技术的增速正在加快，技术的力量正在以指数级的速度扩张……在几十年内，以信息为基础的技术将涵盖所有人类的知识和技能领域，最终包括人类大脑自身的模式识别能力、解决问题的技能，以及情感和道德。

尽管库兹韦尔对奇点的定义相当宽泛，但这个术语已经逐渐被当作一个特定概念：奇点是指计算机智能(通用意义上)超过人类智能的一个假想点。有人认为，到达奇点以后，计算机可以开始运用自己的智能来改进自己，这个过程就会持续自我完善。之后，这些改进后的机器会用它们改进后的智能进一步改进自身，以此类推。从这一点上说，自奇点之后，仅仅依靠人类的智慧就不可能重新获得计算机的控制权了。

这个想法听起来很有道理，也十分令人恐惧。但我们先暂停一下，看看奇点背后的逻辑。简言之，库兹韦尔的推论主要基于这样的观点：计算机硬件(处理器和内存)的发展速度很快会超过人脑的信息处理能力。他的推论引用了计算机领域一个著名的定律——摩尔定律。摩尔定律是计算机处理器公司英特尔的联合创始人戈登·摩尔(Gordon Moore)在20世纪60年代中期提出的，所以以他的名字命名。晶体管是计算机处理器的基础组成单元，在芯片上安装的晶体管数量越多，芯片在给定时间内能够处理的工作量就越大。摩尔定律指出，半导体固定面积上的晶体管数量大约每隔18个月就会翻一番。简单地说，按照摩尔定律，计算机

处理器的功率每隔18个月就会提升一倍。摩尔定律有几个重要的推论，其中之一是计算机处理器的能耗会以同样的速度降低，处理器本身的体积也会逐渐缩小。近50年来，摩尔定律一直被证明是非常可靠的，而在2010年前后，现用处理器技术开始触及物理极限。

现在来看，库兹韦尔的推论隐晦地将奇点出现的必然性与计算机的运算能力简单地关联起来，然而，这种关联是合理的吗？请允许我做一个思考实验，想象一下，我可以把你的大脑复制到一台电脑上(不用紧张，仅仅是想象)，假设用来装载你的复制大脑的电脑是有史以来运行速度最快、最强大的电脑，有惊人的运算速度，那么你就会拥有超级智慧吗？当然，你可以“快速思考”，但这会让你突变得极其聪明吗？从某种微不足道的意义上来说，我想或许会——但是，从任何有意义的智力层面上来说，不会，更不可能让你突破奇点[95]。换句话说，单纯的计算机处理能力的提高不会导致奇点的必然出现。或许这是一个必要条件(如果没有高性能的计算机，我们无法实现人类级别的人工智能)，但并不是充分条件(仅仅拥有高性能计算机并不能让人工智能实现宏伟梦想)。再换句话说，人工智能软件(例如机器学习)的改进速度比硬件发展速度要慢得多。

怀疑奇点主义还有其他理由[96]，一方面，即使人工智能真的能达到人类级别的智能化，也不意味着它就能够以超出我们理解的速度提升自己。正如本书中已经明确的那样，在过去的60年里，我们的人工智能发展之路极其缓慢——那又有什么证据证明类似人类智力水准的通用人工智能能够迅速提升人工智能的发展速度呢？

也有人论证说人工智能系统互相合作以获得超出人们理解或者控制的智能(参见本章开头提到的脸书事件)。但我同样不认为这种论证具有说服力：假设你聚集了1000个爱因斯坦的克隆体，它们的集体智慧会是爱因斯坦的1000倍吗？事实上，我怀疑它们的集体智慧远远无法达到这个数字。再说一次，虽然1000个爱因斯坦克隆体可以比1个爱因斯坦更迅速地完成一些事情，但并不代表它们就变得更聪明了。

基于这些以及更多的原因，我认识的大多数人工智能研究人员都对奇点主义持怀疑态度——至少在可预见的未来，在计算机和人工智能技术方面，还不知道有哪条路能把我们从现在的位置带到奇点。但一些严肃的评论员仍对此感到担忧，并声称我们无视奇点的观点太过自负，他们认为核能就是最好的反面教材。早在20世纪30年代初，科学家们就知

道有大量的能量被锁在原子核中，却不知道如何释放，甚至不知道能否释放。一些科学家对利用核能的想法嗤之以鼻：卢瑟福(Rutherford)勋爵是彼时最著名的科学家之一，他认为，人类妄想能利用这种能量，简直是“自作聪明”。但是，讽刺的是，就在卢瑟福否定核能可用性的第二天，物理学家利奥·西拉德(Leo　Szilard)在伦敦一边过马路，一边仔细思考卢瑟福的声明，突然就冒出了核连锁反应的点子。10年后，美国向日本城市投放原子弹，释放出恐怖的能量，而这一切，就来源于西拉德的灵光一闪。人工智能会不会有一个利奥·西拉德式的灵光闪现呢？一个突如其来的顿悟，把我们很快带向奇点？当然，我们不能排除这种可能性，但它确实微乎其微。核连锁反应实际上是一个非常简单的机制，连中学生都可以理解它。过去60年人工智能研究的所有经验告诉我们，人类水准的人工智能并非如此。

遥远的未来是否会出现奇点？100年后，或者1000年后？在此，我不得不承认，很难说。试图预测计算机技术在100年后(更别提1000年后了)会是什么状况，真是不明智的行为。但在我看来，如果奇点出现了，就像《终结者》里面的场景那样，这简直不可思议。用罗德尼·布鲁克斯的比喻，把人类的智慧想象成波音747，我想问问，我们有没有可能仅凭灵光一闪就发明波音747？或者在毫无预期的情况下就把波音747发明出来了？显然不可能。当然，也存在质疑的观点，哪怕出现奇点的可能性微乎其微，但一旦它出现，对所有人类而言，不啻一场灭顶之灾，所以现在开始要对奇点进行预先思考和计划，是完全有必要的。

不管奇点是否会出现(很显然，我的观点是不会)，但目前看来，确实有不少人很担忧人工智能的发展，认真考虑是否要对其进行监督。我们是否需要法律——甚至国际公约来控制人工智能的发展，就像我们应用核能一样？不过，我认为引入一般的法律来管理人工智能的使用没有可行性，这就有点像试图通过立法来管理数学应用一样滑稽。

我们得谈谈阿西莫夫

每当我和一个普通观众讨论《终结者》的故事时，总有人建议，我们要在构建人工智能的时候就考虑约束问题，避免它成为无法控制的杀人机器。在这之后不久，通常就有人建议我们考虑著名科幻作家艾萨克·阿西莫夫(Isaac　Asimov)提出的机器人三定律。机器人三定律是由阿西莫夫在机器人系列故事中提出的，在他的故事里，机器人装备了强大的人工智能——“正电子大脑”。阿西莫夫最早在1939年制定了三定律，在

接下来的40年里，机器人三定律为他提供了巧妙的情节设计，最终成就了一系列著名短篇小说，以及一部令人失望的好莱坞电影[97]。阿西莫夫故事中人工智能的“科学性”，即正电子大脑毫无意义，当然，这丝毫不影响故事的趣味性。对我们来说，最有趣的是机器人三定律本身：

> 第一定律：机器人不得伤害人类，也不得因为不作为而让人类受到伤害。
>
> 第二定律：机器人必须服从人给予它的命令，除非该命令与第一定律冲突。
>
> 第三定律：机器人在不违反第一、第二定律的情况下要尽可能保证自己的生存。

真是完美的设定，乍一看巧妙又稳妥地解决了问题。那么，我们可以在构建人工智能的时候内置这些定律吗？

关于阿西莫夫的机器人三定律，首先要说的是，它们设计得虽然精巧，但阿西莫夫的故事很多时候都是发生在定律有缺陷或者互相矛盾的情况下。例如，在故事《环舞》中，一个名叫SPD-13的机器人要无休止地围着一池熔融的硒绕圈，因为要遵守人类下达的采硒命令(第二定律)和保护自己(第三定律)之间存在冲突，所以它只能在离硒矿湖固定的距离处绕圈子，如果它再靠近一点硒矿湖，保护自己的定律开始起作用，让它远离；而它要是离得过远，采硒矿的命令又起作用，它必须服从。在阿西莫夫的故事中还有许多其他例子(我强烈推荐你去读一下)。因此，三定律本身虽然很巧妙，但绝不是无懈可击的。

但机器人三定律更大的问题是，它在人工智能系统中根本无法实施。

想想实施第一定律意味着什么，当人工智能在考虑每一个动作时，都需要考虑这个行为可能产生的影响，大概需要考虑全人类(或者能够涉及的部分人类)以及未来可能出现的影响(总不能只关心现在吧)。这是不可行的。另外，“不得因为不作为而让人类受到伤害”这一条也存在同样的问题：系统能够做到不断地思考它涉及的每个人所有可能的行为，并且思考自己是否有相关行为能够阻止他们受到伤害吗？这也是不可行的。

即使单单捕捉“伤害”这个概念也很困难，试想一下：当你坐飞机从伦敦飞往洛杉矶时，你消耗了大量的自然资源，并在途中产生大量污染和排放大量二氧化碳。毫无疑问，这肯定会对某些人产生伤害，但这种伤害的程度不可能精确量化。如果你不相信的话，我可以告诉你，我认

识一些不愿意坐飞机旅行的人，他们的理由就是如此。所以，我想，一个遵守阿西莫夫三定律的机器人是不会坐飞机的。事实上，我怀疑它不能做任何事情，它可能只会蜷缩在某个角落，躲在世界的一隅，因为优柔寡断而如同废铁。

因此，虽然阿西莫夫的机器人三定律为人工智能系统的构建提供了高层次的原则性指导(事实上，我认为大部分人工智能研究人员都会暗地里认同这些准则)，但是，像阿西莫夫的故事里那样，把三定律真正编码进入人工智能系统中的想法，并不现实。

这样一来，阿西莫夫的三定律，以及其他善意的人工智能伦理准则对我们没多少帮助，那么我们又该如何考虑人工智能系统行为的可接受性呢？

我们还是别谈电车难题

阿西莫夫机器人三定律可以说是第一次也是最著名的尝试，试图构建能够管理人工智能决策系统的总体框架。不过它并不是一个严肃的人工智能伦理框架，我们可以把它视为伴随人工智能飞速发展而产生的一系列类似框架的鼻祖。人工智能伦理框架已经成为一项重要的研究课题[98]，在本章剩余部分，我们将深入探索这项工作，并讨论它是否朝着正确的方向发展。我们的探索从一个特定的场景开始，它很有名，吸引了很多人的注意。

电车难题是伦理哲学领域最著名的思想实验之一，最初由英国哲学家菲利帕·富特(Philippa Foot)于20世纪60年代末提出[99]。她引入电车难题的目的是解开围绕堕胎道德的某些高度感性的问题。富特的电车难题有许多版本，最常见的版本是这样的(参见图21)：

一辆电车失去控制，正高速冲向五个无法移动的人。轨道旁边有一个操纵杆，如果拉动操纵杆，电车将转向另一条轨道，那里只有一个人(同样无法移动)。你如果拉动了操纵杆，你会杀死一个人，但会拯救五个人。

那么，你拉还是不拉？

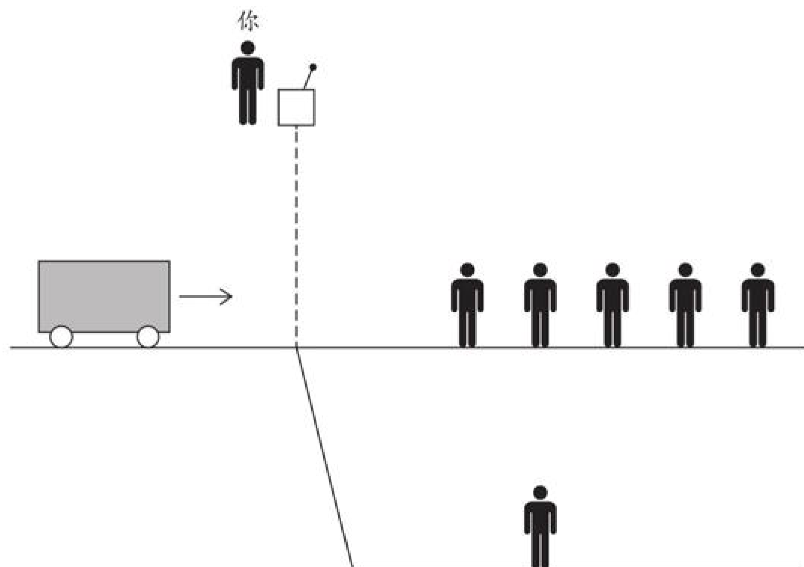


图21　电车难题

如果你无作为，上方轨道的五个人会死；如果你拉动操纵杆，下方轨道的一个人会死。你该怎么办呢？

由于无人驾驶汽车即将到来，电车难题迅速凸显。专家们指出，无人驾驶汽车很可能陷入类似电车难题的困境，然后人工智能软件就会被要求做出艰难的选择。2016年，一则标题为“自动驾驶汽车已经在决定杀死谁了”[100]的网络头条新闻，在网上掀起了一场轩然大波。我认识的几位哲学家受宠若惊地发现，突然有一批网友关注他们，期待他们对这个迄今为止在伦理哲学方面很难下定论的问题发表意见。

电车难题表面看很简单，但它衍生出一系列令人惊讶的复杂问题。我对电车难题的直觉反应是，在不考虑其他条件的情况下，我会拉动操纵杆，因为只死一个人总比死五个好。哲学家称这种想法为结果主义者(因为它根据行为的后果来评估行为的道德性)。最著名的结果主义应该是功利主义了，它起源于18世纪英国哲学家杰里米·边沁(Jeremy Bentham)和他的学生约翰·斯图亚特·密尔(John Stuart Mill)。他们提出了一个被称为“最大幸福原则”的理论，大致来说就是一个人选择的任何行为，都会使“世界总体幸福”最大化。在更现代的术语中，我们会说功利主义者是为了社会福利最大化而行动的人。在这里，功利被定义为社

会福利。

虽然总体原则没问题，但要精确定义“世界总体幸福”并不容易。例如，假设电车难题中的五个人是邪恶的杀人凶手，而另一个人是无辜的小孩。五个邪恶的杀人凶手生命的价值会超过一个无辜小孩吗？如果不是五个，是十个邪恶杀手呢——这会让你下定决心拉动操纵杆吗？

另一种观点则认为，如果一项行为符合普世的“善意”行为原则，那么它是可以接受的，标准例子就是“夺取别人性命是错误的”原则。如果你坚持这个原则，那么任何导致别人死亡的行为都是不可接受的。因此，信奉这样原则的人不会对电车难题采取任何行动，即无作为，因为他们的行为会导致谋杀，虽然不采取任何行动也会导致人死亡。

第三种观点基于美德伦理学的思想，从这个角度看，我们认为一个“有道德的人”体现了我们渴望决策者身上有的美德，然后，我们可以得出结论，他在这种情况下所做的选择，就是正确的选择。

当然，对人工智能而言，做决策的是一个智能体——它必须决定一辆无人驾驶汽车在直行杀死五个人和转弯杀死一个人之间如何选择。那么，当人工智能遇见电车难题或者类似问题时，智能体应该怎么做？

首先，我们应该扪心自问，在这样的情况下，期望人工智能做出正确选择是否合理。如果世界上最伟大的哲学家都无法彻底解决电车难题，那么期待人工智能系统去解决它，合理吗？

其次，我得指出，我开了几十年车了，从来没遇见过这样的难题，我认识的所有人也没有遇见过。此外，我所知道的道德伦理，尤其是有关电车难题的伦理，仅仅是上述你读到的假设情况而已，我不需要通过道德伦理考试才能拿到驾照。到目前为止，这个问题还没有对我造成任何困扰，驾驶汽车也不需要更深层次的伦理推演。因此，要求无人驾驶汽车在上路之前先解决电车难题，在我看来有点荒谬。

第三，在这类伦理问题上，无论你认为自己的答案多么理所当然，其他人总会有不同答案，他们也认为那才是顺理成章。麻省理工学院的研究人员已经证明了这个事实，他们为此做了一个聪明的在线实验，研究人员建立了一个名为道德机器的网站，用户可以在网站上看到一系列的电车难题，并被问及如果无人驾驶汽车遇见这样的情况应该怎么做[101]。那些无辜的受害者可能包括男性、女性、肥胖人士、儿童、罪

犯、流浪汉、医生、运动员和老年人，还可能包括猫狗等动物。这项实验引起了网友们的广泛关注，研究人员从233个国家的用户那里收集到了大约4000万份个人决策数据。

这些数据揭示了全球对电车难题中伦理决策的不同态度。研究人员发现了三个关键的“道德集群”，每一个都体现了具有独特特征的伦理框架。研究人员将这些集群命名为西部、东部和南部。西部集群包括北美和大多数欧洲国家；东部集群包括许多远东国家，如日本和中国，以及伊斯兰国家；南部集群包括中美洲、南美洲和拉丁美洲国家。与西部集群相比，东部集群更倾向于保护合法的人而不是罪犯，更倾向于保护行人而不是车上的乘客，而且更倾向于保护年轻人。南部的集群似乎更关心的是如何保护地位高的个体，以及年轻人和女性。研究人员深入研究发现，还有其他的决策预测因素：例如，如果一个国家拥有繁荣的文化或强大的法治，这两种社会特质在预测偏好方面都会起到明显的作用。

麻省理工学院的研究人员将研究结果(人们认为无人驾驶汽车在有轨电车问题的情况下应该如何运行)与2017年德国联邦政府制定的一些有关汽车道德决策的实际指南进行了比较[102]。德国的指南提出了20条建议，例如：

·在危险情况下，必须始终优先考虑拯救人的生命，而不是防止财产损失。

·如果发生事故不可避免，在决定如何行动时，不允许考虑一个人任何的生理特征(年龄、性别等)。

·必须始终弄清楚，目前是人类还是计算机负责驾驶。

·汽车必须记录下任何时间在驾驶汽车的对象。

其中一些指导原则与道德机器实验中所得出的数据相悖：例如，禁止任何基于个人特征的歧视的规定与道德机器所揭示的拯救年轻人的国际偏好形成鲜明对比。试想一下，如果一辆无人驾驶汽车遵循德国的指导方针，不对事故对象做区分，结果导致一名儿童而不是一名老年晚期癌症患者被撞死，那会引起多大的公愤。我举的例子很极端，在此深表歉意，但你应该能明白我的意思。

虽然道德机器实验挺有意思，但我认为它本身和它所涉及的电车难题都无法给我们太多有关无人驾驶汽车人工智能软件的启示。我不相信未来几十年内我们的无人驾驶汽车会遇见这种道德困境。那么，当遇上电车难题的时候，一辆真正的无人驾驶汽车在实际操作中会怎么做呢？

那些致力于研究无人驾驶汽车技术的人对细节也不太清楚，但以我过去几十年在人工智能领域积累的经验来看，最基本的工程原理是最大限度提高预期安全性(换言之，就是最小化预期风险)。这不涉及深层次道德推断的问题——如果真的需要推断，也可能不会比你面对多个障碍物时避开较大的那些需要做的推断更复杂。坦率地说，即使是这种层面的推断也并不是一定会发生的。事实上最有可能的结果是汽车会紧急刹车。也许，在实践中，在同样的情况下，我们也应该尽全力去做到这一点。

人工智能伦理研究的兴起

别作恶。

——谷歌公司座右铭，2000—2015年

还有更多更广泛的问题涉及人工智能和伦理学，这些问题比电车难题更具紧迫性，也更相关。在撰写本书的时候，人们正为这些问题进行激烈争辩。似乎每一家科技公司都想证明他们的人工智能比其他公司的更具有道德性，几乎每星期各家都有新闻稿宣布一项新的人工智能伦理倡议。我们需要回顾一下这些问题，以及思考它们在人工智能未来发展中可能扮演的角色。

最早也是最有影响力的人工智能道德框架之一是阿西洛玛人工智能准则，它是由一群人工智能科学家和评论员于2015年和2017年在加州的阿西洛玛度假胜地确定的。阿西洛玛人工智能准则一共有23条，全世界人工智能科学家和开发者都被要求签署这些准则[103]。大多数准则都是没有什么争议的，比如：第一条，人工智能研究的目标应该是创造有益的智能；第六条，人工智能系统应该是安全可靠的；第十二条，人们应该有权访问、管理和控制与人工智能系统相关的数据。

另外一些准则志存高远。例如：第十五条要求“人工智能创造的经济繁荣应该被广泛分享、造福全人类”。我个人对签署这一条倒是没什么异议，但这一条对那些大型企业而言，恐怕只是嘴上说说罢了，指望他们落实这一条那恐怕是天真过了头。大企业投资人工智能的主要原因是希望获取竞争上的优势，为股东带来利益，而不是他们想造福全人类[104]。

最后的五条原则涉及人工智能长远的未来，以及对人工智能可能会以某种方式失控的担忧——《终结者》场景再现：

·能力警惕：我们应该避免关于未来人工智能能力上限的过高假设，但这一点还没有达成共识。

·重要性：高级人工智能能够代表地球生命历史的深远变化，人类应该报以高度关注，以及用合适的资源来进行规划和管理。

·风险性：必须投入与人工智能系统预期影响力相对应的努力来应对和缓解它们带来的风险，尤其是毁灭性风险或者涉及人类存亡的风险。

·递归式自我改善：对于设计中能够进行递归式自我改进(自动提高它们的智力，然后利用改进后的智力进一步提高自己)或者自我复制，可能会导致人工智能智力快速提升或者复制品迅速增加的人工智能系统，必须遵守严格的安全和控制措施。

·共同利益：超级智能的开发应当只为普世认同的道德理想服务，应当惠及全人类，而非惠及某一国家或者组织。

再次重申，我个人对签署以上准则是没有任何异议的，但事实上，正如我之前讲到过的，我们离这些准则暗示的场景还有十万八千里，将这些场景纳入准则中几乎就等于制造恐慌。用人工智能科学家吴恩达A的话来说，现在担心这些问题就像在担心火A吴恩达，1976年生，华裔美国人，斯坦福大学计算机科学系和电子工程系副教授、人工智能实验室主任。吴恩达是人工智能和机器学习领域最权威的学者之一，也是在线教育平台Coursera的联合创始人。星人口过剩一样[105]。或许在未来的某一天，这些问题足以将人折磨到失眠，但现在就提出来，是在误导人工智能的未来发展之路，更令人担忧的是，它分散了我们的关注。我们究竟该关注哪些问题，这个会在下一章讨论。当然，也正是因为这些场景在未来很长一段时间内都不可能出现，所以各大公司都可以不费吹灰之力地欣然接受，并且享受由此带来的提升企业形象的正面宣传效果。

2018年，谷歌发布了自己的人工智能道德指南。比阿西洛玛准则略简洁，它们涵盖了许多相同的领域(有益、避免偏见、安全)。并且，谷歌还就人工智能和机器学习开发的最佳实践提供了一些具体指导[106]。2018年底，欧盟提出了另一个框架[107]，还有一个框架是由IEEE(电气和电子工程师协会，一个非常重要的计算机和信息技术专业学会)提出的[108]。许多大公司(不仅仅是IT公司)也发布了他们自己的人工智能道德准则。

当然，大企业宣称他们致力于发展人工智能道德是一件好事情。然而，他们是否真正理解所承诺的东西，这才是难点。高层的愿景是很美好的，比如分享人工智能的益处，肯定受欢迎，但是将其转化为具体行动却并不容易。谷歌十多年来使用的公司座右铭是“别作恶”，这听起来

不错，我敢说这是真诚并带有善意的——但这对谷歌的员工而言又意味着什么呢？如果需要防止谷歌越轨到黑暗面，他们还需要更具体详细的指导。

在已经提出的各种框架内，某些主题反复出现，围绕它们所达成的共识也越来越多。我在瑞典于默奥大学的同事维吉尼亚·迪格努姆(Virginia Dignum)将这些问题分为三个关键类别：解释义务、责任和透明度[109]。

解释义务主要是指，比如一个人工智能系统做了一个对某人有重大影响的决策，那么这个人有权要求系统对这个决策进行解释。但是怎么样才算是解释，这就是个难题了，在不同的环境下有不同的答案，而现在的机器学习程序无法提供解释。

责任则意味着应该明确对决策负责的智能体，而且，最重要的是，我们不应该试图声称人工智能系统本身要对决策“负责”，责任方应该是部署该系统的个人或者组织。这就指向了一个更深层次的哲学问题，一个与道德智能体有关的问题。

道德智能体通常被理解为一个实体，它能够分辨是非，并理解其行为所导致的后果。人们通常认为，人工智能系统可以承担道德智能体的责任，并且能够为它的决策和行为负责。而人工智能研究界的普遍观点恰好相反：软件是不能被追究责任的。更进一步说，人工智能研究中的责任并不意味着制造出有道德责任的机器人，而是以负责任的方式开发人工智能系统。例如，发布一个类似Siri的软件智能体，误导用户以为自己在跟另一个人交互，这就是软件智能体的开发者对人工智能不负责任的使用。软件在这里不是罪魁祸首，开发和部署它的人才是。负责任的设计在这里意味着人工智能将始终清晰地表明它的非人类身份。

最后，透明度意味着一个系统使用的数据应该是可获取的，其中使用的算法也应该是清晰明确的。

人工智能伦理研究的兴起是令人值得高兴的进步，尽管目前正在提出的各种框架和体系实际的实施范围还有待观察。

谨慎地表达意愿

有关人工智能伦理的讨论有时候会让我们遗忘一个平凡的现实：人

工智能软件就只是软件而已，我们不需要创造什么新奇的技术让软件出错。简言之，软件本身就有缺陷，没有缺陷的软件是不存在的：只是有的软件因为缺陷崩溃了，而有的没有。开发无缺陷软件是计算机领域的一项重要研究，发现和消除缺陷是软件开发的主要内容之一。但是人工智能软件为引入缺陷提供了新的方式。其中最重要的一点是，如果人工智能软件要代替我们工作，我们需要告诉它希望它做什么，这往往不像想象中那么容易。

大约15年前，我正在研究一种技术，旨在使车辆在不需要人为干预的情况下进行自我协调。听起来很酷炫，不过因为我研究的特定场景是铁路网，所以实际情况相对要简单一些。铁路网是环形网络，上面有两辆列车朝着相反的方向行驶。当然，火车和铁路都是虚拟的——没有实际的轨道(事实上连玩具轨道都没有)。假设虚拟的铁路通过一个狭窄的隧道，如果两辆火车同时进入隧道，那么就会发生(虚拟的)车祸，而我的目标是阻止这一切。我尝试开发一个通用框架，允许我向系统提出一个目标(本例中的目标是防止火车撞车)，系统将返回一些规则，列车如果遵循这些规则就能保证目标实现(列车不会发生碰撞)。

我的系统开始工作了，但跟我想象的差距甚远。当我第一次向系统输入目标时，系统返回的规则是：两列火车必须都保持静止。当然，这是可行的——如果两列火车都保持静止，当然不会发生车祸了，可这不是我想要的方案。

我遇见的问题是人工智能研究中的典型问题，实际上在计算机科学中也存在。我们想把自己的意愿传递给计算机，这样计算机可以代表我们去达成它。但是，将意愿准确地传达给计算机，本身就是一个非常有问题的过程，原因有好几个。

首先，我们可能并不知道自己想要什么，至少并非明确知道，在这种情况下，表达自己的意愿几乎不可能。另外，我们的意愿通常存在矛盾，在这种情况下，人工智能又要如何理解它？

此外，我们不可能一次说清自己的偏好，所以通常我们所做的是对意愿和偏好进行概述，而概述和全面的叙述之间总会存在差距，人工智能又该如何弥合这些差距呢？

最后，也许也是最重要的一点，当我们和人类交流的时候，通常默

认彼此间有共同的价值体系和规范。我们不需要每次互动之时都把所有的东西交代清楚。但人工智能并不清楚这些默认的价值体系和规范，它们必须得到明确的说明，或者我们需要通过某种方式保证人工智能系统的后台存在这些东西。如果没有，那我们没法得到自己想要的结果。在上文的火车铁轨研究中，我传达了我的目标，即火车要避免撞车，但我忘了传达一个信息：火车仍然需要保持运行。如果是跟人交流，我想所有人都会理解并默认这一点，哪怕我忘记交代。但计算机系统不会。

牛津大学哲学家尼克·博斯特罗姆(Nick Bostrom)在他2014年出版的畅销书《超级智能》[110]中讲述了这种情况，他称之为不通情理的实例化：计算机按照你的要求去做了，但并没有按照你预期的方式。想象不通情理的实例化具体案例，可以让人不停地想上几个小时：你要求机器人确保你的房子不会被窃贼入侵，它索性一把火把房子烧了；你要求机器人保证人类不会得癌症，它干脆把所有人都杀了。诸如此类。

当然，我们在日常生活中也经常遇见这类问题：每当有人设计了一套旨在鼓励某一类行为的激励机制时，总有人会找到某种博弈方式，在不按预期行事的情况下获得奖励。我想起了苏联时期的一则逸事(可能是编造的)：苏联政府希望鼓励刀具生产，因此决定根据刀具的重量来奖励生产刀具的工厂，结果如何？餐具工厂很快开始生产重得拿不起来的餐刀餐叉之类……

迪士尼经典电影的影迷可能会想到一个相关的情景，1940年迪士尼电影《幻想曲》中有一段情节，天真的巫术学徒米老鼠厌倦了从井里打水并提到屋里的家务活儿，为了减轻自己的负担，他召唤了巫术扫帚来做这件事。但是当米老鼠打瞌睡醒来后，他不得不阻止扫帚一桶又一桶地往屋里提水，结果他的地下室被水淹没了。最终他不得不寻求巫师师父的介入来纠正这个问题。米奇的扫帚完成了他的要求，但那并不是他想要的。

博斯特罗姆还设想了以下场景：假设有一套控制回形针生产的人工智能系统，人们要求它“最大化生产回形针”，然后，从字面意思来讲，系统将考虑先把地球和宇宙的其他部分转化成回形针的样式。同样，这个问题归根结底还是沟通问题：在这种情况下，我们传达目标的时候，要确保明确无误，不会产生歧义。

解决这个问题的方法是设计一种人工智能系统，以尽量减少其行为

对周围环境的影响。也就是说，我们希望人工智能实现目标，同时让它所涉及的一切都尽可能保持或接近现在的状况。**“ceteris paribus preferences”**(即“尽可能保持其他条件不变”)的概念说明了这一点[111]。“ceteris paribus”是拉丁文，意思是“其他条件不变”。因此，按照“尽可能保持其他条件不变”的想法，如果我们让人工智能系统做一些事情，是希望它完成任务的同时，保持其他一切尽可能不发生变化。因此，当我们发出“避免我的房子被盗贼入侵”指示时，我们的意思是“避免我的房子被盗贼入侵，同时尽可能使房子的其他一切保持现状”。

解决这些问题的核心都是让计算机理解我们真正想要的是什么。逆向强化学习就是针对这一问题展开的，我们在第五章了解了常规的强化学习：智能体在某种环境中行动，并获得奖励。强化学习的目的是找到一个行动过程，最大限度地获取奖励。在逆向强化学习中，我们首先确定了“理想”的行为(即人类会怎么做)，然后再制定人工智能软件能获得的相关奖励[112]。简言之，我们是将人类的行为视为理想行为的典范。

第八章 现实中的人工智能会导致什么问题

虽然我对出现奇点的可能性持怀疑态度，但这并不意味着我认为人工智能安全无虞。毕竟，人工智能是一门通用技术：它的应用仅仅受限于我们的想象。所有的技术都可能产生意想不到的效果，未来几十年甚至几百年内都存在可能性。而且所有的技术都可能被滥用。我们的无名氏祖先率先用上了火，我们不可能因为他没预料到燃烧化石燃料将导致气候变化而怪罪于他。英国科学家迈克尔·法拉第(Michael Faraday)于1831年发明了电动机，他也没预料到会产生电椅这种刑具。1886年获得汽车专利的卡尔·本茨(Karl Benz)肯定无法预言，他的发明在未来的一个世纪里会造成数百万人的死亡。互联网的发明者温顿·瑟夫(Vint Cerf)可能也没想到恐怖组织会利用他的发明分享可怕的杀人视频。

就像这些技术一样，人工智能也会产生负面的影响，甚至可能会波及全球。像这些技术一样，人工智能也有可能被滥用。这些后果或许不会像上一章我们提到的试图消灭人类的《终结者》场景那么富有戏剧性，但却是我们以及我们的后代在未来几十年中需要解决的问题。

因此，在本章中，我们会讨论一些我认为在人工智能领域真正值得人们担忧的东西，以及它们对未来的影响。最后，我希望你们会认同，人工智能领域可能会出现问题，但不是像好莱坞影片中那样的问题。

我们将从就业和失业开始：人工智能将取代我们工作，这值得我们思考人工智能的存在和发展将如何改变人们的工作结构，以及人工智能担任的职能和人类之间存在异化的可能性。这反过来促使我们思考人工智能技术的应用对人权的影响，还有致命性自主武器出现的可能性。然后我们将考虑算法偏见的问题，以及人工智能缺乏多样性的问题，还有假新闻和假人工智能现象。

就业和失业

机器人会取代我们的工作，我们最好从现在开始早做打算，以免为时已晚。

——《卫报》，2018年

除了《终结者》式的毁灭场景，有关人工智能最广泛和令人恐惧的问题就是它将如何影响未来的职业结构，特别是它拥有使人失业的潜

力。电脑不会疲惫、不会宿醉、不会吵架、不会抱怨、不会要求成立工会，也许最重要的是，它们不要工资。这就是它会让雇主开心、雇员紧张的原因。

尽管类似“人工智能会让你变成无用阶层”的头条新闻在过去几年里层出不穷，但这种恐慌并不是史上第一次出现。人类劳动的大规模自动化至少可以追溯到1760年至1840年之间发生的工业革命。工业革命代表商品制造方式的彻底改变，从小规模生产经营(家庭手工业)转向我们今天熟悉的大规模工厂生产。

工业革命并不是由单一的技术突破推动的，而是一系列的技术进步，加上独特的历史和地理环境促成。英国通常被视为工业革命的发源地，大量本地和从美洲等殖民地运送来的棉花，集中在英格兰北部的工厂进行加工。大英帝国不仅盛产棉花原料，还提供了庞大的成品市场，这为英国纺织工业的繁荣提供了经济条件。从17世纪30年代开始，纺织品加工方面的一系列技术革新使得人们可以制造出更大型、更快捷的纺织加工机器，它能够稳定制造出高质量的纺织品，并且达到家庭手工业无法望其项背的规模。原材料通过利物浦和曼彻斯特大量进口，被带到新兴发展的工业城镇。最初，纺织厂采用水动力，但煤动力蒸汽机的改良让这项技术迅速得到广泛应用，使工厂不再需要依赖普及性和稳定性都不好的水动力源。而因地制宜的英格兰北部正好有着丰富的煤矿藏，为发动机提供燃料，煤矿城镇和工业城镇并驾齐驱，成为工业革命发展的有力支撑。

工业革命开创的工厂形式，使大多数人的工作性质发生了根本性的变化。工业革命前，大多数人直接服务于农业，工业革命把人们从农村带到了工厂所在的城镇，把他们的工作地点从田地和农舍搬到了厂房。他们的工作性质也变成了到今天仍然广泛存在的类型：在工厂从事一项专业化、高度重复性的工作，作为涉及高度自动化生产流水线的一部分。

家庭手工业无法在价格、质量和商品一致性上与新技术竞争，手工业者只能另谋高就(许多劳动力流入了新工厂)，否则就会面临失业。社会发生了巨大的变动，但并不是所有人都喜欢这种改变。在19世纪初期，卢德派运动兴起，卢德派是组织松散的团体，他们反抗工业化，焚烧和粉碎了工厂的机器——他们认为这些机器剥夺了他们的生计。但那只是昙花一现的运动，很快就被英国政府扼杀。在1812年，英国政府将

破坏机器定为可判死刑的重罪。

当然，1760年至1840年发生的只是第一次工业革命，自那以后，技术一直在稳步发展，卢德派反抗工业化的两个世纪以来，发生过好几次类似的巨大变化。具有讽刺意味(并且悲哀)的是，在第一次工业革命中繁荣起来的工业城镇，在20世纪70年代末80年代初又被另一次工业革命摧毁。这一时期传统产业衰落的原因有很多，国际经济全球化意味着，在欧洲和北美传统工业地区制造的商品，现在正被世界另一边新兴经济体的商品所取代——因为更加低廉的生产成本。美国和英国倡导的自由市场经济政策对传统制造业十分不利，但是，最关键的技术“贡献”应该是微处理器的发展，它让工业里面大量非技术性工作得以全面自动化。

微处理器发展是今天计算机技术得以实现的根本，它是一种系统部件，将计算机中央处理器的所有关键部件组合成单一、小型、便宜的单元。在微处理器出现之前，计算机体积庞大，价格昂贵，运行不稳定。微处理器使得计算机变得更便宜、更迅速、更可靠，而且体积缩小了许多，这让自动化生产成为可能，特别是制造业。但英格兰北部的工业城镇因此遭受了致命打击——40年前遭遇制造业寒冬后，它们至今仍未恢复元气。

虽然自动化生产让英国北部传统制造业中心遭受灭顶之灾，但我们必须明白，微处理器等新技术的发展总体是让经济活动净值增长的。新技术创造了新机遇，创造了更多的企业、更多的服务和更多的财富。微处理器发展的情况也确实如此：新兴技术创造的就业机会和财富比毁掉的多，只是不在英格兰北部而已。

劳动自动化也不是什么新鲜事了，但现在令人担忧的事情，就像以前自动化和机械化夺走了非技术工作一样，人工智能可能会从我们手上夺走技术工作的机会。如果人工智能广泛应用，留给我们人类的机会还有多少？

人工智能会改变现有的工作结构和性质，这一点是毋庸置疑的。我们只是不清楚，这种改变是否会像工业革命那样，具有戏剧性和颠覆性，还是会以更温和的方式进行。

2013年，我在牛津大学的两位同事卡尔·弗雷(Carl Frey)和迈克尔·奥

斯本(Michael Osborne)撰写了题为“就业的未来”的报告〈[113]，引发了对这个问题的争论。报告令人震惊地预测：在不久的将来，美国会有高达47%的工作岗位受到人工智能和相关自动化技术的影响。

弗雷和奥斯本将702种职业按照他们所认为的自动化程度进行分类。报告指出，被取代风险最高的职业包括电话销售员、下水道疏通工、保险业务员、数据录入员(实际上还有许多类似职业)、电话接线员、销售员、雕刻工和出纳员。有一定风险的职业包括治疗师、牙医、顾问、医生(包括外科医生)以及教师。他们的结论是：

我们的模型预测，大多数从事运输和物流行业的工人，连同大部分办公室和行政后勤人员，以及生产性职业的劳动者，都处于危险之中。

弗雷和奥斯本还总结出三个不会轻易被人工智能取代工作的特点。

首先，也许是理所当然的，需要创造性思维的工作被认为是安全的。创造性思维职业包括艺术类、媒体类和科学类。

其次，需要较强社交能力的工作也是比较安全的。那些需要理解和管理人际互动和微妙人际关系的工作，不会轻易被人工智能取代。

最后，他们指出，涉及灵敏感知和灵巧手工的工作也很难被自动化。在这里，人工智能面临的问题是，虽然机器感知领域一直在进步，但在这方面和人类相比还差得远。我们可以快速理解高度复杂化和非结构化的环境，而机器人只能应付规则的、结构化的环境，一旦超出机器学习训练的范畴，就会面临重重困难。另外，人类的手还是比机械手灵活得多——2018年，罗德尼·布鲁克斯预测，恐怕要到20年后才能出现一只跟人类一样灵活的机械手[114]。至少在那之前，那些从事需要大量灵巧手工的人是安全的——木匠、电工或者管子工今晚可以睡个踏实觉了。

弗雷和奥斯本的报告受到了广泛的质疑和批评，许多评论员指出了报告中他们所认为的方法论缺陷和不合理假设，并强烈表示这个结果太过笼统。虽然我也认为报告中最悲观的预测在一段时间内不会成为现实，但我坚信人工智能以及先进的自动化、机器人等相关技术将在不久的将来让很多人失业。如果你的工作是分析数据然后做出决策(比如是否提供贷款)，那我不敢说人工智能是否会取代你的角色。但如果你的工作只是用程序化的脚本跟客户对话(就像许多呼叫中心接线员的工作)，那

么我很抱歉，人工智能会很快让你变成无用阶层。如果你的工作只是在一个受限制的、地图清晰的城市地区进行日常驾驶，那么我也很抱歉，人工智能注定会取代你，只不过别问我那是什么时候。

然而，对于我们大多数人来说，新技术主要改变的是我们的工作性质。大多数人不会被人工智能系统取代，反之，使用人工智能工具可以让我们在工作中变得更高效、更优秀。毕竟，拖拉机的发明没有取代农夫，只是让他们成为效率更高的农夫；文字处理机的发明也没有取代秘书，只是让他们成为效率更高的秘书而已。我们可以通过软件智能体来简化工作，提升效率，这样我们可以从无穷无尽的文书表格工作中解脱出来。人工智能将嵌入日常工作的所有工具中，给我们带来难以想象的益处——甚至有很多是以无声无息的方式。用吴恩达的话来说，“如果一项脑力工作，普通人只需要不到一秒钟就能完成，那么，我们可以用人工智能实现它，无论是现在还是将来[115]”。这可是一长串任务列表。

虽然许多人担心人工智能会夺走人类的工作机会，带来全球失业和加速不平等的严重后果。但技术乐观主义者希望人工智能、机器人技术和先进的自动化技术将引领我们走向一个截然不同的未来。这些乌托邦主义者相信人工智能最终会将全人类从枯燥乏味的工作中解脱出来——在未来，所有工作(或者说，至少所有肮脏、危险、无聊或其他不受欢迎的工作)都将由机器完成。人类可以自由自在地写小说、讨论柏拉图、四处游玩或做其他事情。这是一个诱人的梦想，尽管它也是老调重弹——乌托邦式的人工智能在科幻小说中出现得可真不少。(不过，更有趣的是，科幻小说中更频繁出现的是反乌托邦式的人工智能——也许把背景放在人人幸福、人人健康、人人快乐的未来中，要写出一个有意思的故事难度会更大吧。)

在这一点上，反思一下20世纪70到80年代出现的微处理器技术发展是很有必要的。正如我们前文讨论过的，微处理器的出现引发了大规模的工厂自动化浪潮，当时微处理器技术在社会各界引发了各种争论，跟现在人工智能引起的争论颇为相似。彼时，乐观主义者们预测，在不远的将来，我们用来工作的时间将大大减少，每周工作三天或者类似的工作制度将成为常态，休闲时间会大大增加。

只可惜事实并非如此。为什么呢？有一种观点认为，与其享受计算机、自动化和其他技术进步为我们腾出来的休闲时间，不如把时间更多地用在努力工作上，赚更多的钱，享受更好的商品和服务。比如我们想

买彩电、录像机、CD播放机、DVD播放机、家用电脑、手机，我们也想去海外度假，想要更大更快的汽车、更华丽的衣服、更精美的食物，为了支付这一切，我们必须付出更多的努力。有人认为，只有全社会都没有过多欲望，都甘于接受简单的生活方式，没有太多消费品，只需要维持基本生活，我们的工作时间才可能大大缩减。另一种观点认为，技术进步带来的经济利益并没有被平均分配：富人变得更富有，贫富差距越来越大。不管确切的原因是什么，从20世纪70年代的经历中，我们能够得出一个明显而略让人觉得沮丧的结论，至少在不久的将来，技术可能不会为所有人创造一个悠闲的乌托邦。

这就巧妙地为我们引入了全民基本收入的概念：即社会上的每个人都应该获得一定的保障性收入，不管他们是否工作，也不需要任何考核或者测试。全民基本收入并不是一个新出现的概念，但最近的技术发展，特别是人工智能的发展，使它重新成为人们关注的焦点。有一种设想，人工智能/机器人/自动化技术将创造足够的财富，使得全民基本收入成为可能(因为机器可以完成工作)，以及必要(因为，在这种设想下，人们可能没有工作了：机器把人类的工作机会夺走了)。

我也很想相信乌托邦式的未来，但由人工智能推动的大规模全民基本收入计划在短期内实现似乎并不合理[116]。首先，人工智能所产生的经济效益必须是巨大的，才能使全民基本收入得以实现，因此就需要远远超过以往的技术创新来支撑庞大的项目，而没有任何迹象表明目前的人工智能发展会带来如此规模的经济效益。其次，要想推行全民基本收入计划，将需要前所未有的强烈政治意愿：可能会在社会环境极其紧张的情况下才能迫使政府接受这样的行动方针。我的猜测是，恐怕需要出现大规模的失业潮，这个想法才有可能被提上议程。最后，全民基本收入将从根本上破坏社会的结构，毕竟在当今社会中，工作占据着核心的社会地位。目前同样没有任何迹象表明社会做好了接受这种变化的准备。

虽然人工智能是导致工作环境变化的一个重要因素，但它不是唯一因素，甚至可能并不是最重要的。一方面，无情的全球化进程尚未走到终点，在此之前，它将继续以我们无法想象的方式震撼整个世界。计算机本身也没有达到进化的终点——在可预见的未来，计算机将会继续变得越来越便宜、小巧，以及拥有更高的交互性能。光是这些发展就足以持续改变我们的世界，还有生活与工作的方式。在这个背景下，我们的

世界相互联系日益紧密，全世界的社会、经济和政治格局也在不断变化：化石燃料资源减少、气候变化和民粹主义政治抬头。所有的因素都将对就业和社会产生影响，对我们产生影响，而这些影响比人工智能大得多。

算法异化以及改变工作性质

在一本人工智能科普书里看到卡尔·马克思(Karl Marx)的名字时，你可能会感到惊讶，但是，这位《共产党宣言》的合著者所关注的问题，与当前人工智能如何改变工作和社会现状的争论密切相关。特别在19世纪中叶第一次工业革命以后建立起来的异化理论——当时资本主义社会才初具形态。马克思的异化理论关注的是工人与工作和社会的关系，以及资本主义制度是如何影响这种关系的。他指出，工业革命出现了工厂系统，工人们在工厂中从事重复性的、无聊的、最终毫无成就感的工作，以换取微薄的收入。他们没有机会筹备安排自己的工作，事实上，他们根本没有控制工作的权利。马克思认为，这是一种非自然的生活方式。工人们的工作，从真正的意义上来说是没有意义的，但是工人们别无选择，只能从事这种毫无成就感、没有意义的工作。

可能这些工作状态也会让你点头认同，不过人工智能和相关技术的迅速发展将马克思的理论推向了一个新的维度：在未来，我们的老板可能只是一段算法。

我们又一次看到了异化现象，特别是在所谓“零工经济”领域——这是工作模式发生改变的另一个方面。半个世纪前，你或许会在一家公司从毕业干到退休，这很常见。长期雇佣关系是一种常态，那些频繁跳槽的员工会受到质疑。

但现在，长期雇佣关系已经越来越少见，取而代之的是短期工作、计件工作和临时工作——这就是零工经济。在过去20年里，移动计算机技术的发展是全球化零工经济迅速腾飞的主要原因，工作人员可以通过移动设备协调和参与工作。手机内置的GPS系统可以随时定位和监控工作人员的位置，还可以随时向他们传达工作指示。人们日常工作所做的每一件事，甚至在键盘上按键的次数，发送电子邮件的习惯语气，都可以被电脑程序记录和监控。

网络购物巨头亚马逊因为对员工要求严格、监控严密、工作环境严

苛而时常遭受批评，以下是2013年英国《金融时报》的一篇报道，描述了一名仓库工人的工作[117]：

> 亚马逊的软件会计算出最高效的配送货物步行线路，指导工人装满手推车，然后通过手持卫星导航设备屏幕的指示，指导工人从一个货架走到另一个货架。即使路线如此高效，依然有许多需要步行的地方……“你就像个机器一样，不过是人形的，”亚马逊的工人说，“如果你乐意的话，可以管这叫作人类自动化。”

如果马克思今天还活着，我想，他会用这篇报道作为异化概念的完美解释例证。这就是人工智能造就的自动化噩梦：人类的劳动被系统地简化为那些不能被机器或软件自动化的任务，工人被细致地监控和监督着，没有创造力、没有创新性、没有个性，甚至没有思考的空间。想象一下，如果你的年度考核是由电脑程序来决定，甚至是否决定录用你，也是由电脑程序来判断，你会有什么感受。有点愤世嫉俗地说，我们也许不应该太关心这些问题，因为这些工作不会存在很长时间了：很快它们就会被人工智能和机器人自动化技术完全替代(亚马逊正在大力投资仓库机器人技术)。

世界上很大一部分劳动人口从事这样的工作，这个前景一点都不诱人。不过，我想现在已经说得很清楚了，这并不是什么新鲜事——只是从工业革命时期就存在的趋势，人工智能只是扩展了它的维度。当然，当今职场上还有许多人的工作条件远比亚马逊的仓库工人差得多。这不应该归罪于技术，技术是中立的。雇主、政府、工会和监管者应该思考人工智能在这些方面对社会的影响——它将如何改变人类工作的性质，以及什么样才算体面的工作环境和条件，还有失业问题。迄今为止，人工智能所引起的焦虑和争议，大多集中在失业问题上。

人权

上述讨论显示了控制我们工作状况的人工智能系统是如何让现代职场人异化的，但是，人们对人工智能的使用的更大担忧，是它影响了最基本的人权。拥有一个人工智能系统做老板已经很让人头疼了，它告诉你什么时候可以休息，什么时候应该工作，给你设定目标，并对你的工作状况随时监控和批评……但是，如果人工智能系统有权决定你是否应该进监狱，这种情况又会如何？

再一次声明，这可不是臆想——现在已经有类似的人工智能系统正在投入使用。例如，英国杜伦的警察部队于2017年5月宣布，他们将启

用危害评估风险工具(Harm Assessment Risk Tool， HART)[118]，这是旨在帮助警察决定应该释放还是拘留涉案嫌犯的人工智能系统。HART是机器学习的经典应用，它的训练库包括2008至2013年获得的所有拘捕数据(约104 000起拘捕事件)，并用2013年的全年拘捕数据进行测试(测试用的数据未纳入训练库)[119]。最终结果是，在低风险案例中该系统的准确率高达98%，而在高风险案例中为88%。有趣的是，系统设计的时候考虑到了谨慎处理高风险案例(例如涉嫌暴力犯罪的案件)，或者这能解释最终结果的区别。系统使用34种不同的案例特征进行训练，其中大部分与嫌犯的犯罪史有关，不过也包括年龄、性别和住址特征。

没有迹象表明，杜伦的警察把所有决定嫌犯是否拘留的决定权都交给了HART，该系统只是作为决策支持工具，辅助监管人员做决策而已。但是，尽管如此，当社会各界得知HART投入使用时，仍然引起了很大的不安。

主要问题在于HART只关心独立案件的一系列特征，它不会像一个有经验的监管人员那样，会综合考虑嫌犯背景和案件流程，所以它的重要决定是基于相当狭窄的数据基础做出的。决策缺乏透明度也是一个令人担忧的问题(这是机器学习的经典问题，人工智能程序无法给出决策的解释)。当然，培训所使用的数据特征可能存在偏差，也有人提出这一点(尤其是将嫌疑人的住址作为特征之一，这更令人担忧，该系统的决策可能涉嫌歧视底层区域的人)。

还有一点，尽管HART只是一个决策支持工具，但可能在未来的某一天，我们会发现它成了主要的决策者：监管人员有可能太过疲惫、太过迷糊或者太过懒惰，放弃了自己来做决策的权利，于是不加任何思考地全盘接受HART的建议。

我认为，所有这些担忧都指向一点：HART这样的系统侵蚀了人类判断的地位。归根结底，对一个人做出重大影响的决策者是人类，比起别的东西，更让人感觉自在。毕竟，我们享有基本的人权，例如接受同侪审判的权利，而这项权利来之不易，理应受到高度重视。让类似HART的电脑程序对人类做出判决，也许会让人感觉我们在放弃这些来之不易的权利。对这样的趋势我们不应当掉以轻心。

这些担忧是合理的，尽管我不认为应该简单粗暴地全面禁用类似HART的系统，但在使用它的时候，应当遵循非常重要的注意事项。

首先，诚如HART案例中明确指出的，这些工具只能用于为人类决策者提供支持，而不能取代人类进行决策。机器学习决策并非十全十美，它们难免会做出一些明显毫无意义的决策。而当前机器学习还有一个令人沮丧的现状，就是很难确认它什么时候的决策会出错。因此，在给人们带来重大影响的决策方面，盲目遵循人工智能的建议是非常不明智的。

另一个问题涉及对机器学习技术不成熟地开发和利用。HART是由经验丰富的研究团队开发的，他们应该全面周到地考虑过这款软件可能面临的各类问题。并不是所有开发者都有如此丰富的经验，或者会如此深入思考。因此，令人担忧的是，代替人类做决策的系统，是否在开发环节考虑得足够谨慎和全面。

HART只是备受人权组织关注的应用于执法系统的软件之一，伦敦警察厅因使用一款名为“帮派矩阵”的工具而饱受批评。该系统有数千人的记录，并将数据代入一些简单的数学公式中，以预测这些人参加帮派的可能性[120]。帮派矩阵系统大部分似乎都建立在传统计算机技术上，不太涉及人工智能领域，但有这方面的趋势。国际特赦组织称该系统是一个“种族偏见数据库，将某个年代的黑人全部定罪”。有人声称，如果你表现出只偏好某一类音乐的倾向，就会被列入数据库。美国一家名为PredPol的公司销售支持“预测警务”的软件，用以预测犯罪热点区域[121]。同样，该软件的使用引发了人权方面的争议：如果数据有偏差怎么办？如果软件设计不合理怎么办？如果警察开始过度依赖它该怎么办？另一个系统，COMPAS，旨在预测一个人再次犯罪的可能性[122]，该系统用于量刑参考[123]。

有一个极端的案例说明这种错误的想法是如何走向绝境的。2016年，两名研究人员发表了一篇文章，声称只需要让人工智能系统识别一下人脸就可以发现潜在的犯罪行为。这样的系统让我们回到了一个世纪前那些武断的犯罪理论。随后的研究表明，他们的系统大概是根据一个人是否微笑来判断犯罪行为：系统训练使用的是警方拍摄的罪犯面部照片，而这些照片往往都不会面带微笑[124]。

机器人杀手

希望你会认同我的想法，我认为人工智能系统当我们的老板会导致员工异化，而让人工智能系统来决定一个人是否应该被警方拘留，这是

对人权的一种侮辱。那么，对于一个有能力决定你生死的人工智能系统，人们又将如何看待呢？随着人工智能的迅猛发展，人们对这一前景的担忧开始见诸报端，其中至少有一部分原因是我们提到过很多次的《终结者》场景。

关于自主武器的话题总是带有强烈的煽动性，许多人会本能地厌恶它，认为这样的系统是不道德的，永远不该被建立。对于持有这样观点的人而言，若是有品行端正的人持有相反观点，会让他们感到惊讶。因为我知道这个主题有多么震撼，所以我会尽可能轻描淡写地带领你深入研究潜在的人工智能自主武器所引发的问题。

自主武器的讨论大多源于战争中使用得越来越频繁的无人机。无人机，顾名思义，就是无人驾驶的飞机，在军用领域，它可以携带导弹等武器。因为无人机不用搭载人类飞行员，所以它比传统飞机更小巧、更轻便、造价更低。而且无人机的飞行不会给遥控它的人带来风险，因此它可以被用于飞往危险区域执行任务。对于军事组织来说，这自然是一个富有吸引力的选择。

在过去的50年里，无人机的军用开发经历了数次尝试，不过直到21世纪才成为现实。自2001年以来，美国一直在对阿富汗、巴基斯坦和也门的军事行动中使用军用无人机。尽管我们不知道具体数量，但似乎美国已经投入了数百架无人机，可能导致了数千人死亡。

遥控无人机本身就引发了各种严重的伦理争议，例如，由于控制无人机的驾驶员没有亲身涉险，他们可能会采取一些实际在场的情况下不会采取的行为。而且，最重要的是，人们不会像自己亲临现场那样认真对待行为的后果。

出于以上以及更多的原因，使用遥控无人机引起了极大争议。然而，自主无人机出现的可能性将人们的担忧提升到了更高层面。无人机不再被遥控，意味着很大程度上，它们不再需要人类的引导或者说干预来执行任务。而且，作为执行任务的一部分，它们可能有权决定是否夺走人类的性命。

自主无人机和其他种类自主武器的构想，立刻会让你联想到非常熟悉的电影场景：机器人大军无情地、精准地使用致命武器屠杀人类，没有任何对人类的怜悯、同情或理解。我们已经了解到为什么这样的《终

结者》式末日场景不值得我们今晚就担忧得失眠，但是自主武器出问题的可能性，以及致命的后果，仍然是人们反对其发展和应用的一个重要论据。但是，关注自主武器也有许多其他原因。其中之一是拥有自主武器的国家可能不太介意发动战争，因为他们的部队不需要亲自上前线。因此，自主武器可能使得发动战争的决定变得更加容易，使得战争更加普遍化。但是，最常见的反对意见是，自主武器是不道德的：人们不能制造能够决定夺走人类性命的武器，这是错误的。

值得指出的是，在目前的技术条件下，人工智能驱动自主武器是完全可能实现的。请考虑如下场景[125]：

现在有一种廉价的小型无人机型，配备了摄像头和GPS导航系统，机载小型电脑和一块手榴弹大小的炸药。这架无人机被设计成可以在城市道路巡航，寻找人类。它不需要精准识别特定的人，只需要识别人类的形态，当它识别出人体时，会飞到人类的位置，引爆炸药。

这种设想十分可怕，人工智能只需要能够在街道上空导航、识别人体并飞向他们就可以了。我认为，任何一个合格的人工智能研究生，只要有合适的资源，就能够构造一个这样的无人机原型，而且它们可以以低廉的价格大量生产。现在想象一下，若是成千上万的无人机被投放到伦敦、巴黎、纽约、德里或者东京街头，将会发生什么。想象一下大屠杀，想象一下恐怖的画面。据我所知，这样的无人机已经有人制造出来了[126]。

对许多人而言，关于人工智能自主武器是否应该存在，根本无须争议，然而，有关的争议确实存在。

这场争辩中最著名的主角之一，是美国佐治亚理工学院的罗恩·阿尔金(Ron Arkin)教授，他认为自主武器的出现肯定是不可避免的(总会有某个人，或者某个机构将它们最终制造出来)，因此，最好的对策是思考如何设计它们，使它们的行为比普通人类士兵更合乎道德[127]。毕竟，他指出，人类士兵在道德方面并没有很好的记录。当然，他也承认，不可能出现拥有“完美道德”的人工智能自主武器，尽管如此，阿尔金教授仍然相信，我们能够制造出比人类士兵更有道德的自主武器。

还有其他支持自主武器的观点，例如，有人认为，让机器人去从事卑鄙的战争事业，总比让人类去打仗更好：战争的赢家往往是拥有更优

秀机器人的一方(当然，问题就变成了如何确保我们拥有更好的机器人)。

我还听过这样的论点，即反对自主武器比反对常规战争更没有道德。例如，当一架B-52轰炸机在5万英尺的高空飞行，释放炸弹的时候，负责释放炸弹的投弹手并不清楚他们投下的32吨的炸弹将落到何处，也不知道它会落到谁身上。那么，为什么人们要反对能够精准杀人的自主武器，而不反对这样随意杀戮的常规武器轰炸？我想，答案是我们应该同时反对它们。但实际上，常规轰炸确实没有像自主武器那样引起这么多的道德争议。

不管人们试图构建什么有利的论据，我必须坚定地说，国际人工智能界的大多数研究人员都强烈反对开发致命的自主武器。关于自主武器可能被设计成比人类士兵更合乎道德的论点并没有被广泛接受。理论上这是一个有建设性的想法，但实际上我们根本不知道怎么去做，而且也不可能很快成为现实(见上一章的讨论)。虽然我敢说，提出这一论点的初衷是最好的，但令人担心的是，它已被那些想现在就制造自主武器的人所劫持。

在过去十年里，科学家、人权组织和其他组织联合开展了一场旨在阻止自主武器研发的运动。“禁止机器人杀手运动”成立于2012年，得到了国际特赦组织等主要国际人权组织的支持[128]。该运动的既定目标是实现全球禁止研发、生产和使用完全自主武器。这项运动得到了人工智能研究界大部分人的支持。2015年7月，近4000名人工智能科学家和2万多人签署了一封公开信，支持禁令；除此之外还有许多其他类似的倡议。有迹象表明，政府组织听到了人们的呼声。2018年4月，中国驻联合国代表团提议禁止使用致命的自主武器；英国政府最近的一份报告也明确建议，永远不要给人工智能系统以伤害或杀人的权力[129]。

显然，致命的自主武器既危险又不道德，因此，在我看来，按照《渥太华禁雷公约》[3]的思路，实施一项禁令是有必要的。但这可能比较难：不仅要考虑到致力于制造这类武器的组织会施加阻力，另外，制定一项禁令本身不是容易的事情。英国上议院人工智能特别委员会在2018年的报告中指出，给致命自主武器下一个清晰的定义相当困难，这将是立法的一个主要障碍。从纯粹实际的角度来看，仅仅试图在武器开发领域禁用人工智能技术基本上不太现实(我斗胆猜测许多常规武器已经在某种形式上使用人工智能系统控制)。此外，立法禁止某种特定的人工

智能技术使用，比如神经网络，那就更不现实了，因为软件开发人员可以很轻易地在代码中掩饰使用的技术。

因此，即使公众和政府都愿意控制或禁止发展和使用致命自主武器，制定和执行相关立法也是困难重重。但是，有迹象表明，各国政府正在尝试，这未尝不是一个好的开始。

算法偏见

我们也许希望人工智能系统能够做到公正和公平，摆脱困扰人类世界良久的偏见和成见，但恐怕事实并非如此。在过去十年里，随着机器学习系统被推广到越来越多的应用领域，我们开始了解到自动决策系统也会表现出算法偏见。现在它已经成为一个主流研究领域，许多团体都在努力理解算法偏见带来的问题，以及如何避免。

算法偏见，顾名思义，是指计算机程序——不仅仅是人工智能系统，而是任意计算机程序——在其决策过程中表现出某种形式的偏见的情况。该领域的主要研究者之一凯特·克劳福德(Kate Crawford)指出，存在算法偏见的程序可能会造成两方面的伤害问题[130]。

首先是分配伤害，分配伤害体现在某个群体在某些资源方面被拒绝(或优待)的时候。例如，银行可能会使用人工智能系统来预测潜在客户的优劣——优质客户意味着会按时还贷，而劣质客户则更可能拖欠贷款。他们可以使用优质客户和劣质客户的相关资料来训练人工智能系统，经过训练的系统可以查看潜在客户的详细信息，并预测客户可能是优质还是劣质客户，这是一个经典的机器学习程序的案例。但如果程序存在偏见，那么它可能会拒绝向某个群体提供贷款，或者偏袒某个群体，更可能放款给他们。在这里，偏见导致了相关群体明显可识别的经济障碍(或优待)。

另一种是代表性伤害，通常发生在系统产生强化刻板印象或者偏见时，一个最臭名昭著的例子就是2015年，谷歌的照片分类系统将黑人照片标注为“大猩猩”[131]，强化了令人反感的种族刻板印象。

但是，正如我们在第一章里了解到的那样，计算机不过是执行指令的机器而已，它怎么会存在偏见呢？

因为机器最重要也是单一的获取信息途径是数据，偏见就是通过数

据引入的。机器学习程序使用数据进行训练，如果数据本身就存在偏差，那么程序也将学习数据中隐含的偏见，而训练数据本身就可能存在不同程度的偏差。

最简单的可能性是，构建数据集的人本身就是带有偏见的。在这种情况下，他们会在数据集中嵌入这类偏见，可能不会太明确，也可能是无意识的。事实上，无论每个人认为自己多么公平公正，我们都存在某种形式的偏见。这些偏见将不可避免地在我们创建的训练数据中体现出来。

虽然这些问题是人为的，但是机器学习也会不知不觉地成为帮凶。例如，如果机器学习程序的训练数据不具有代表性，那么该程序的决策将不可避免地被扭曲。例如，假设银行根据一个地理区域的数据集对贷款软件进行了训练，那么这个程序很可能最终会对其他地区的个人产生偏见。

设计拙劣的程序也会产生偏见，举个例子，假设在上面提到的银行案例中，你选择用种族血统作为关键特征来训练银行的人工智能系统，最终，你的程序毫无疑问会在衡量发放贷款的时候带有明显的种族歧视。(你不会真的觉得银行能蠢到这个地步吧？等着看呗。)

算法偏见是目前存在的一个突出问题，我们之前就了解过，当前人工智能系统大多采用机器学习的方式构建，机器学习的特点是“黑盒”：它无法用人类能够理解的方式解释或者说明它的决策。如果我们过于信任人工智能系统，这个问题就会更加严重——有迹象表明已经存在这样的趋势了。银行建立了人工智能系统，使用包含几千个案例的数据集训练它，训练结果是看上去它能够得出和人类专家相同的结果，所以人们就武断地相信人工智能系统不会出错，并且毫不犹豫地信任它的决策，而不进行深入思考。

有时候，你会感觉到似乎世界上的每一家公司都在疯狂地将机器学习应用到他们的业务中，但在这种疯狂浪潮中，他们有可能创造出更多带有偏见的程序，因为他们根本没意识到这其中的致命问题，最关键的是，要用正确的方式获取数据。

多样性缺乏

1956年，当约翰·麦卡锡向洛克菲勒研究所提交了关于人工智能达特

茅斯暑期学校的提案时，他列出了拟邀请参加会议的47人名单。“并非所有人都能来参加这次会议，”他在1996年写道，“只有那些我认为会对人工智能产生兴趣的人，才会被列入名单。”

现在，我想提个问题，你猜一下有多少名女性被列入了邀请名单，见证了人工智能作为一门新兴科学的诞生呢？

答案是：零。我非常怀疑，任何一个有声望的当代研究资助机构会考虑赞助这样一场连基本的多样性测试都通不过的活动。事实上，现在的标准做法是要求申请人确保如何解决平等性和多样性问题。但是，在当时男女就是这么不平等。如果你一直关注本书，很难忽略的一个事实就是，人工智能长期以来似乎都是一个由男性主导的学科。

事后来看，我们可以意识到当时普遍存在的不平等现象，并对此感到遗憾。我同样认为，用我们今天还没有彻底实现的标准去评判20世纪中叶举行的一次活动也没多大意义。更重要的问题在于，今天的人工智能研究本质是否有所不同。在这里，虽然有一些好消息，但总体情况只能说喜忧参半。一方面，你不管参加哪一个有声望的国际人工智能会议，都会看到许多女性研究人员的身影。但是，另一方面，人工智能领域的男性主导现象仍然存在：缺乏多样性仍然是人工智能现阶段面临的一个棘手问题，就如科学和工程学科诸多领域一样[132]。

人工智能研究领域的性别构成很重要，原因有很多。一方面，男性主导的研究领域将使潜在的女性科学家感到反感，从而扼杀了这一领域潜在的宝贵人才。而且，也许更重要的是，如果人工智能完全是由男性设计的，那么我们最终得到的结果将是——找不到更好的术语来形容了：男性人工智能。我的意思是，这样建立的体系必然会体现出特定的世界观，而这种世界观不能够代表女性，甚至对女性根本不友好。如果你不相信，那么请你读一本让我对这个问题大开眼界的书：《看不见的女人》，卡罗琳·克里亚多·佩雷斯(Caroline Criado Perez)所著[133]。她提出的关键问题在于：我们这个世界上几乎所有的东西都是由只考虑单一性别的人设计和制造的——男性。她认为，造成这种情况的根本原因是她所说的“数据缺口”：通常用于制造和设计的历史数据集绝大多数是以男性为导向的：

> 大部分历史记录的数据都是有缺陷的……古时候狩猎都是由男性来完成的，以往的历史写作者根本没有给女性留下多少空间……相反，人们默认男人的生活方式代表了人类的生活方式，当谈到人类另一半的生活时，通常只有沉默和空白……这些沉默，这些缺陷，都会产

生后果。它们每天都影响着女性的生活。这种影响都发生在细微处。例如，办公室空调温度标准是为男性设定的，让女性在办公室里瑟瑟发抖；或者，女性在超市里挣扎着想够到按照男性身高标准设计的货架顶层等……它们不危及生命，不像一辆汽车的安全设施没有考虑到女性的身体尺寸，不像心脏病发作的时候得不到救治和诊断，所以你的这些烦恼都是“非典型的”。这样的现状，即女性每天都生活在围绕着男性的数据构建的世界中，带来的后果可能是致命的。

卡罗琳以各种毁灭性的细节全面记录了以男性为基准设计和以男性数据为标准的问题到底有多普遍。当然，对于人工智能而言，数据是必不可少的，但数据集里男性偏见是普遍存在的。有时候，这种偏见很明显，比如在广泛应用于训练语音理解的口语数据集中，69%的数据集是男性声音，不可避免地，这个语音理解的系统对女性声音的理解就会比对男性声音的理解效果差得多。但是，有时候，这种偏见更细微。如果你收集一组厨房的照片来训练程序，那么这些照片大多会跟女性有关；或者假设你收集一组主要描绘各大公司CEO的照片，会发现大部分都是男性。现在，你应该可以很轻易推断出这样的数据会让程序造成什么样的偏见。而且，正如卡罗琳指出的，这些事情都确确实实发生过。

有时候，这种偏见简直是根深蒂固。一个臭名昭著的案例发生在2017年，有人发现谷歌翻译公司在翻译文本的过程中，有时候会篡改文本中的性别[134]。如果你将下列文字从英语翻译成土耳其语：

他是一名护士

她是一名医生

再将翻译好的土耳其语翻译成英语，就变成了：

她是一名护士

他是一名医生

还真是没有性别的刻板印象啊，是吧？

偏见本来就是人工智能中存在的问题，对女性而言，这个问题更具有特殊性。因为新的人工智能系统建立在有着强烈男性偏见的数据基础上，而建造新人工智能系统的团队压根没注意到这一点，因为他们自己也有男性偏见。

人工智能面临的终极挑战之一，就是如果我们构建某些体现真人特性的系统，例如，当我们构建Siri和Cortana这样的软件智能体时，可能会无意中以强化某种刻板的性别认知方式来构建。如果我们要构建一个具有服从性的人工智能系统，让它看上去或者听上去都像女人，那么就

等于在无意中宣传女性的仆从地位[135]。

假新闻

顾名思义，假新闻就是将虚假的、不准确的或者误导性的消息当作事实，并以新闻的方式呈现。当然，在数字时代之前，世界上就存在大量的假新闻源，但是互联网，尤其是影响巨大的社交媒体，却成了传播假新闻的完美渠道。

社交媒体的存在是让人们可以方便地彼此联系，现代社交媒体平台在这方面取得了惊人的成功——西方的脸书和推特，还有中国的微信，都有着大量的用户基础。21世纪初，社交媒体应用首次出现的时候，它们似乎都是关于朋友、家人和日常生活的：有很多小孩子的照片，还有猫狗等宠物的。但社交媒体作为一个非常强大的工具，很快就被应用于其他目的。互联网假新闻频发的现象已经在2016年登上了头条，这一现象证明了社交媒体有多强大，以及多么轻易就能造成严重影响。

有两件事情让假新闻成为热点：2016年11月美国总统大选，唐纳德·特朗普(Donald Trump)最终被选为美国总统；2016年6月，英国就是否继续留在欧盟举行全民公投，结果支持退出欧盟的一方以微弱优势胜出。毫无疑问，社交媒体在这两次投票中都扮演了重要角色——无论人们如何看待他的政治及个人行为，特朗普是公认的善于利用社交媒体召集支持者的人。在这两个案例中，都有人认为推特等社交媒体平台被用来传播有利于最终胜利方的假新闻。

人工智能是假新闻的重要组成部分，因为它对假新闻的传播方式至关重要。所有的社交媒体平台都依赖于你在它们上面消磨了多少时间——这就给了平台向你展示广告的机会，而这也是平台最终盈利的方式。如果你喜欢在平台上看到的内容，那么你就会花更多的时间在它上面。因此，社交媒体平台就有动机向你展现你喜欢的东西，而不是真相。它们怎么知道你喜欢什么呢？因为你会告诉它们。每一次你的点赞，都会让平台去查找类似的消息，即你可能会喜欢的消息。一个运行良好的平台都会建立一张有关你偏好的关系图(这就是残酷的事实)，并以此来决定给你推送什么内容[136]。例如，约翰·多伊(John Doe)是一名白人男性，喜欢暴力视频，有种族主义倾向……如果某个平台想要约翰·多伊喜欢它，那么，它会向这位用户推送什么样的新闻呢？

人工智能在其中所起的作用是根据你的点赞、评论，以及你点击的相关阅读链接等行为去分析你的偏好，然后寻找到你也许会喜欢的新消息。这些都是很典型的人工智能应用，所有社交媒体平台公司都会有专门的研究和开发团队来解决这些问题。

但是，如果所有社交媒体平台都这样——弄清楚你喜欢什么，向你展示你喜欢的东西，隐藏你不喜欢的——那么，它们呈现在约翰·多伊面前的是怎么样的世界？一个充满了暴力视频和种族主义新闻的世界，他不会得到多渠道的、公正的信息，他将深陷于自己的社交气泡中，对自己的世界也会存在偏见，并且偏见还会因此得到加强：这被称为证实性偏见。而这种现象如此令人担忧的原因是它的规模：社交媒体正在全球范围内操纵人们的信仰，无论这种操纵是蓄意的，还是纯粹偶然，这无疑是值得警惕的。

从长远来看，我们看待世界的根本角度有可能被人工智能改变。每个人都通过自己的感官(视觉、听觉、触觉、嗅觉和味觉)获取关于世界的信息，并利用这些信息建立对现实的一致看法：对世界实际情况的描述和可以被广泛接受的看法。如果你和我都以同样的方式见证了一个特定的事件，那么我们会得到同样的信息，我们可以利用这个信息表达对共识现实的一致看法。但是，假如这个世界没有了共识现实，没有共同看法，会发生什么呢？如果我们每个人都以完全不同的方式看待世界，又会发生什么呢？人工智能可能会让这些假设成为现实。

2013年，谷歌发布了一款可穿戴设备——谷歌眼镜，它的外观和日常佩戴的眼镜类似，只是谷歌眼镜配备了一个摄像头和一个小型投影仪，并可以通过蓝牙与智能手机连接。发布之后，人们立即担心谷歌眼镜中“隐藏”的摄像头会被用于拍摄一些不恰当的照片，但该设备真正的潜力在于内置投影仪，它可以将用户看到的任何内容与投影图像叠加。这其中的应用潜力是无穷的。例如，我个人最不擅长的一件事情是认人：有些人我明明很熟，可就是叫不出名字，这是一件非常尴尬的事情。一款谷歌眼镜的应用程序就可以帮助我识别出眼前的这个人，并且谨慎地提醒我他的名字。

这种类型的应用被作为增强现实(Augmented Reality， AR)应用：它们获取真实的世界，并在其上覆盖计算机生成的信息或图像。但是，如果应用程序不仅仅是增强现实，而是完全改变现实，甚至使用户根本察觉不到这种改变，又会如何？在撰写本书的时候，我的儿子汤姆已经

13岁了，他是托尔金(Tolkien)的著作《指环王》的超级粉丝，也特别热衷于这本巨著衍生的电影。想象一下，谷歌眼镜应用程序可能会改变他对学校的认知，让他的同学看上去像是精灵，老师看上去像是兽人。你可能也想要一款类似的星球大战主题的应用程序，等等。这会很有趣，但问题是：如果我们都居住在自己的私人世界里，那共识现实又有什么意义呢？你和我不再有相同的、共通的经验来建立共识。另外，这样的应用程序也容易被黑客入侵。想象一下，如果有一天汤姆的眼镜被黑了，从根本上改变了他认知世界的方式，从而操纵他的信仰，那会发生多么恐怖的事情。

尽管目前这样的应用还不能实现，但在未来二三十年内，这类增强现实应用很有可能成为现实。我们已经开发出了人工智能系统，它能够生成对人类而言完全真实的图像，但其实图像是由神经网络构建的。例如，在撰写本书时，人们就表示了对DeepFakes的担忧[137]。DeepFakes是一款人工智能换脸工具，可以通过换脸技术，让图片或者视频中出现根本没有出现过的人。负面影响最大的案例发生在2019年，美国众议院议长南希·佩洛西(Nancy Pelosi)的一段讲话被篡改，让人觉得她有语言障碍，或者可能受到毒品或酒精的影响[138]。DeepFakes也被用来修改色情视频，在视频中加入实际上没有参与的“演员”[139]。

目前，DeepFakes修改过的视频质量还很差，但它正在慢慢变好，或许很快我们就完全无法分辨照片或者视频是真的，还是DeepFakes篡改过的。到那时，照片和视频将不再是对某一事件的原始可靠记录。如果我们每个人都居住在自己特有的、由人工智能驱动的数字世界里，那么建立在共同价值观和原则之上的社会将面临真正的危险。社交媒体上的假新闻，只是一个开始。

人工智能造假

我们在第四章中看到了类似Siri、Alexa和Cortana这样的软件智能体是如何在21世纪开始的十年里，作为20世纪90年代智能体研究的直系“后裔”出现的。在Siri问世后不久，一些大众媒体就报道过Siri存在未经记录的功能。你可以对Siri说：“你好，Siri，你是我最好的朋友。”Siri会给出一个看上去很有意义的回应(虽然我尝试的结果是“好的，迈克，不管天气好坏，我们都会成为好朋友的”)。媒体们很兴奋，他们想知道，这是通用人工智能吗？好吧，不是。Siri所做的只是针对某些关键词和语句在常备回答中选择回应，这和几十年前的ELIZA没有

太大差别。在这个问答里面没有什么智能可言，显然，这个有意义的答案不过是假的人工智能而已。

苹果并不是唯一一家使用假人工智能的公司，2018年10月，英国宣布一款名为Pepper的机器人将参加下议院听证会[140]，但这只是无稽之谈。Pepper机器人当然出席了听证会(而且，根据记录，它是一个很了不起的机器人，是许多幕后精彩研究的结晶)，但它只是用预先写好的答案回答预先确定的问题而已，甚至还不到ELIZA的水平。

在Siri和Pepper的案例中，没有任何欺骗的意图——它们只是一点善意的小乐趣。但许多人工智能研究界人士对此感到很恼火，因为这些小乐趣描绘了一幅完全错误的人工智能远景图。首先，如果有人真的相信Pepper可以回答人们提出的问题，他们就会对目前人工智能问答系统产生不切实际的期望，这完全是误导性的。不过我想应该不会有太多人上当吧——毕竟很容易就能看出问题。那么问题就在于，人们可能会认为这些胡说八道作秀才是人工智能的真正目的：造假。

类似的事件发生在2018年12月，在俄罗斯雅罗斯拉夫尔的青年科学论坛上，一个“高科技机器人”表演了令人震惊的舞蹈。结果证明，这是一个穿着机器人服装的真人[141]。目前尚不清楚活动的组织者是蓄意欺骗还是作秀，我姑且认为这只是一个善意的玩笑。但是俄罗斯国家电视台报道称，这个“机器人”看上去跟真人一样。任何观看了这场舞蹈的人都会得出一个合理的结论，这应该能代表目前机器人技术和人工智能所能达到的最高水准。

不幸的是，这样的造假在人工智能的边缘领域太普遍了，这让人工智能研究人员非常沮丧。2017年10月，沙特阿拉伯宣布将公民身份授予一个名叫索菲亚的机器人，这毫无意外地引起了媒体的广泛关注[142]。鉴于沙特阿拉伯在人权方面不太理想的记录，尤其是在妇女权益方面，许多评论员强调了这个宣布(赋予机器人公民身份)的讽刺意味。但许多人也想知道，这个机器人是否存在一些全新的东西：它是通往通用人工智能的起点吗？

索菲亚由汉森公司制造，是一款仿真机器人，外形和人类一样，还能拥有丰富的面部表情，当然也拥有用自然语言交谈的能力。商业内幕网站于2017年12月发布了对索菲亚的“采访”[143](索菲亚的回答使用斜体)：

你的感受和喜好是什么？

你有没有让机器人住进你的家里，或者跟你一起工作过？

没有。

你知道，或许现在已经有不少机器人进入你的生活，而你并没有意识到。你希望有一天能够和机器人一起生活吗？

那现在我的身边出现的是什么样的机器人呢？

确实。

当你不知道怎么回答的时候，你默认的答案就是“确实”吗？

是的。

脸书的人工智能研究主管亚恩·勒库恩(Yann Lecun)对此嗤之以鼻。2018年1月，他在推特上写道：“这些把戏对比真正的人工智能，就跟变戏法对比真正的魔法一样天差地别。这些简直就是胡说八道。”这恰好概括了人工智能研究界对这种廉价宣传噱头的不屑。

人工智能造假，主要是人们被误导了，相信他们看到的就是真正的人工智能，而事实上，这些幕后都只是一些把戏而已。我听到有传言说，如果人工智能初创公司不能让他们的技术在关键的成果展示中发挥作用，他们就会在幕后使用假人工智能来糊弄。我不清楚这种做法有多普遍，但人工智能造假对每个人来说都是实实在在的问题，也是对人工智能研究界的莫大伤害。

第九章 通往有意识的机器之路

我希望到目前为止，这本书能够成功地让你明白一件事情，虽然近年来人工智能和深度学习方面取得了真实的、令人兴奋的突破，但它们并不是构建通用人工智能的法宝。深度学习可能是通用人工智能的一个重要组成部分，但它绝不是唯一的组成部分。实际上，我们并不清楚还缺失了哪些关键部分，更不知道通用人工智能的秘方究竟是怎么组成的。我们开发出的所有令人印象深刻的人工智能系统——图像识别、语言翻译、无人驾驶汽车，都无法构成通用人工智能。从这个意义上来讲，我们仍然面临罗德尼·布鲁克斯在20世纪80年代强调的问题：我们有一些智能组件，但不知道如何将它们组成一个真正的通用智能系统。无论如何，某些关键的组件仍然缺失，正如我们在第五章所看到的，即便是当代最好的人工智能系统也无法展示出它们对自己所做的事情有着真正意义上的理解。尽管它们非常擅长自己的工作，但它们仍然只是为了执行特定的、狭隘领域的任务而构建和优化的软件组合而已。

因为我相信，我们离通用人工智能还有十分漫长的道路要走，所以对于强人工智能的目标——构建跟人类一样有自我意识的，真正能够自主存在的机器，我自然是表示怀疑的。不过，这是最后一章了，让我们放纵一下吧，即使强人工智能前景渺茫，我们仍然可以从探索中寻找到乐趣，也仍然可以思考如何朝着它前进。所以，让我们一起沿着通往有意识的机器的道路旅行吧，让我们想象这里的风景是什么样子的，又可能会遭遇哪些障碍。并且，最重要的是，我们如何知道即将接近这条道路的终点。

意识、思想和其他奥秘

1838年，英国科学家约翰·赫歇尔(John Herschel)进行了一个简单的实验，试图测量太阳辐射有多少能量。他把一个装有水的容器暴露在阳光下，测量了太阳能使容器中水温升高1摄氏度所需的时间。通过一个简单的计算，赫歇尔可以估算出我们的恒星每秒发射多少能量。结果令人难以理解：在一秒钟内，太阳辐射出难以想象的能量，这一数量远远超过地球上一年所产生的能量。这就给当时的科学界出了难题：新出现的地质证据表明，我们生活的地方至少有几千万年的历史(所以太阳至少也有这么长的历史)，但还没有已知的物理过程可以为太阳在那么长的

时间里提供能量。任何已知的能源都会导致太阳最多在几千年内就烧毁。当时的科学家们天马行空，发明出令人着迷的难以置信的理论，试图将赫歇尔简单、易于重复的实验证据与地质记录的证据相协调。直到19世纪末，核物理学诞生，科学家才开始了解原子核中潜在的庞大能量。在赫歇尔实验出现整整一个世纪之后，物理学家汉斯·贝特(Hans Bethe)最终提出了目前被广泛接受的关于恒星能量产生的解释，即核聚变[144]。

现在我们回到强人工智能的话题，我们的目标是建造真正具有意识、具有思维，能够拥有自我意识和理解力的机器，与我们自身非常相似。而目前的我们就跟当年的赫歇尔处在同样的位置，因为人类思维和意识这种现象——它们是如何进化的，如何工作的，甚至它们是如何在我们的行为中扮演控制角色的——对我们而言，就像在赫歇尔时代为太阳提供能量的物理机制，是完全神秘的。这些问题我们不知道答案，连寻求答案的方式都不太清楚。目前，我们只有一些线索，以及大量的猜测。事实上，如果这些问题有了明确的、令人满意的答案，我们就能够从科学意义上理解宇宙的起源和命运。正是这种根本性的缺乏使得强人工智能离我们如此遥远——我们都不知道该从什么地方着手。

实际上的情况更加糟糕，因为我们甚至不知道到底该处理什么。我也提到过好多次“意识”“思想”和“自我意识”之类的术语，但事实上，我们都不知道它们具体是什么东西。看上去这些概念都很容易理解——毕竟我们都拥有它们——但我们没有办法用科学的方式来定义或者衡量它们。从科学意义来讲，我们无法证明它们真实存在，但根据个人经验和常识，它们确实存在。赫歇尔能够用一个实验来解决他的问题，使用了很好理解、可以量化的物理概念：温度、能量，等等。我们根本没有这样的测试来研究思想或者意识：它们不适合被客观观察或测量。没有标准的科学单位来衡量思想或者主观意识，甚至没有直接的方法来确定如何衡量它们——我看不出你在想什么，我也无法明白你的感受。从历史上看，我们对人脑结构和运作的大部分了解都是通过研究那些因疾病或创伤而大脑受损的人获得的，但这很难成为一个系统的研究项目。虽然像核磁共振成像之类的神经成像技术给我们提供了大脑结构和运作的重要见解，但它们并不能让我们了解个人的主观体验。

当然，尽管没有精确定义，我们还是可以确定一些在讨论意识时出现的共同特征。

我们需要先确定一个重要的观点，意识产生于有主观感受的智能体，拥有主观上的内在感受性。重点就在于对内在心理现象的感知，哲学上称之为感受性。这个名字挺别致，不过含义其实很简单：感受性是指所有人都会经历的精神感觉，比如咖啡的味道。让我们暂停一下，先想一想这个香味，或者，干脆去给自己煮上一杯，吸入那个香气。你所经历的那种感觉，就是感受性的实例。在炎热的夏天喝一杯冰啤酒的经历，从冬天到春天气候开始转暖的经历，我的孩子们在探索新事物上取得成功让我感觉到开心的经历……这些感受都是定性的：你能够了解它们，并且自己也很享受其中，但矛盾的是，尽管我们谈论的是相同的经历，但我并不确定你的感受是否跟我一样。因为感受性——以及其他心理体验——本质上都是私人的。我闻到咖啡香时候的精神体验只有我自己能理解，我无法判断你是否有过类似经历，即使我们可以用同样的词句来描述它。

1974年，美国哲学家托马斯·内格尔(Thomas Nagel)对意识的争论做出了最著名的贡献[145]，内格尔提出了一种测试方法，通过这个测试，人们可以分辨出某个事物是否具有意识。假设你想知道下方列表中的事物是否是有意识的：

·一个人

·一只猩猩

·一条狗

·一只老鼠

·一条蚯蚓

·一个烤面包机

·一块石头

内格尔提出的测试是对上述实体思考问题“成为一个X是什么样的感觉”。如果我们认为成为一个X是有感受的(这里的X可能是人，可能是猩猩)，那么内格尔就认为，这个智能体X是有意识的。看看上面的列表，这个问题应用到人身上，结论是肯定的。猩猩呢？我想成为一只猩猩应该也是有感受的，狗和老鼠也一样。因此，根据内格尔的测试，猩猩、狗和老鼠都是有意识的。当然，这并非它们确实有意识的“证据”，我们在这个问题上不得不依赖常识而不是客观事实。

蚯蚓呢？在这里，我认为我们可能拿不准了。蚯蚓是一种很简单的生物，根据内格尔的测试，成为一条蚯蚓会有感觉吗？我很怀疑这一

点，因此，根据内格尔的论点，蚯蚓是没有意识的。当然，也有人会认为，蚯蚓也应该有一些简单的意识，但我不得不坚持说，那跟人类的意识比起来，实在是太简单了。不过，在烤面包机和石头的问题上，我不接受任何质疑：显然它们都是无意识的。成为一个烤面包机，可真的是毫无感觉。

内格尔的测试提出了一些重要的论点。

首先，意识并不是一个有或者无的东西，它是有层次的，从极端的成熟的人类意识，到另一个极端的蚯蚓的简单意识。即使是人与人之间，也存在差异，一个人有意识的程度也会不同，取决于他们是否受酒精等外部因素的影响，或者仅仅是因为太过疲惫。

其次，意识对不同的实体而言是不同的。内格尔的论文是《成为一只蝙蝠是怎样的感受》，他选择蝙蝠作为标题，是因为蝙蝠和人类差异非常大。内格尔的测试在蝙蝠这个实体上肯定是有意义的，根据内格尔的理论，蝙蝠具有意识。但是蝙蝠有我们没有的感觉，尤其是声呐，它们在飞行的时候发出超声波，根据超声波的回音来感知周围环境。有些蝙蝠甚至能探测到地球磁场，并利用它进行导航——它们的体内自带罗盘。人类没有这项感知，因此，我们无法想象成为一只蝙蝠的具体感受，尽管在内格尔测试中这个答案是肯定的。蝙蝠的意识与人类的意识完全不同，事实上，内格尔相信它超出了人类能够理解的范畴，尽管我们仍可以确定它的存在。

内格尔研究的主要目的是对意识进行测试(“成为一个X是什么样的感觉”)，并得出结论，有些意识超越了我们能够理解的范畴(我们无法想象成为蝙蝠是什么样的感觉)。但他的测试也可以应用于电脑，大多数人似乎都相信成为一个电脑没什么感觉，就像烤面包机一样。

基于这个原因，有人用内格尔测试的论点来证明了强人工智能是不可能的，因为成为一个计算机是没有感觉的。我个人并不接受这个结论，在我看来，“成为一个X是什么感觉”不过是基于直觉的回答，直觉能够很容易区分显而易见的案例(比如猩猩和烤面包机)，但我不明白，在更微妙的情况下，或者在我们对自然界的体验之外(比如人工智能领域)，我们怎样将直觉当作可靠的指引。也许我们无法想象成为一台计算机是什么感受，仅仅是因为它与我们完全不同，但这并不意味着(至少对我来说)机器意识是不可能存在的——机器意识只是跟人类意识不同而

已。

试图证明强人工智能不可能实现的观点有许多，内格尔的观点只是其中之一，我们再来看看其中最著名的几种。

强人工智能不可能存在吗

内格尔的观点基于一种常识，即人类是有生命的物体，这是人类特殊的地方。而计算机不同于人，它是没有生命的，所以，强人工智能是不可能出现的。根据这一论点，我和老鼠的共同点比我和电脑的共同点更多，电脑和烤面包机的共同点比它和我的更多。

对此我有异议。在我的观念里，不管人类有多么了不起，最终也就是由一群原子组成的物体。人类和人类的大脑都是物理实体，遵守物理定律——即使我们目前并不太清楚这些物理定律是什么。人类是非凡的、奇妙的、不可思议的生物，但从宇宙及其规律来看，我们并没有什么特殊性。当然，这不能回答一个难题，即一堆特定的原子是怎么产生意识的——我们稍后会回到这个问题上。

美国哲学家休伯特·德雷福斯(Hubert Dreyfus)提出了“人类是特殊生命体”的变体理论。德雷福斯在批判人工智能的时候有一个主要观点，人工智能的实际成就实在是配不上人工智能这个名字。在这方面，他并非完全没有道理。但他有一个具体的论点，用以否定强人工智能存在的可能性。即人类的许多行为和决策建立在“直觉”的基础上，他认为直觉不像计算机要求的那样精准。简言之，德雷福斯认为，人类的直觉不能够简化为计算机程序那样的步骤。

现在确实有大量的证据表明我们的许多决策不是基于明确或者严密的推理[146]，我们经常做出决定，但无法阐明自己的理由。事实上，我们大多数决策都是这一类型。从这个意义上来说，我们确实依靠某种直觉。但这种直觉肯定是源自我们随着时间推移获取的经验(要么是通过进化获得的经验，要么是通过基因传递给我们的经验)，即使我们无法在意识层面上表达出来，这也并非什么神秘的事情。而且，正如我们所看到的，计算机可以在经验中学习，并成为有效的决策者，即使它们也无法清楚地表达自己决策的基本原理。

反对强人工智能存在的最著名论断，来自哲学家约翰·希尔勒(Joln Searle)，正是他创造了强人工智能和弱人工智能的术语，我们在第一章

就了解过。他发明了一个名叫“中文房间”的场景，试图证明强人工智能是不可能实现的。中文房间场景如下：

想象一个人被反锁在一个房间里，房门上有一个卡槽，他可以通过卡槽收到卡片，上面写着中文的问题。他自己不懂中文，但是房间里有规则书，可以指导房间里的人如何按照规则用中文写一个答案，然后把它传递出去。此时此刻，这个房间(包括里面的一切)实际上正在进行一个“理解中文”的图灵测试，房间里提供的答案让测试员相信测试对象是一个理解中文的人。

接下来，扪心自问：房间里的人真正懂得中文吗？希尔勒不这么认为，房间里的人不懂中文，房间本身也不懂中文，不管从什么角度来看，整个处理过程都没有表现出对中文的理解。这个人所做的一切不过是精确仔细地按照规则书的步骤给出问题的答案。他的人类智慧在这个场景里面仅仅用在尽职尽责遵守规则回答问题上。

显然这个人所做的就是计算机的工作：简单地按照一系列指令执行一系列步骤。他所执行的“步骤”就是计算机的指令。因此，根据希尔勒的观点，按照同样的道理，即使通过图灵测试的计算机，也没有表现出理解力。

如果希尔勒的观点是正确的，那就意味着理解这种能力——强人工智能所需要的能力——是不能够通过遵循步骤执行命令产生的。因此，用传统计算机是无法实现强人工智能的。如果认可这一点，这个简单的论点将扼杀人工智能的宏伟梦想。不管你的程序看上去拥有多么出色的理解力，这都只是一种错觉：在程序的背后，没有半点理解可言。

对于希尔勒的批评，人们提出了许多反对意见。

一个明显的基于常识的反对意见就是中文房间根本不可能实现。除此之外，让人去扮演计算机处理器的角色，也就意味着要花费上千年的时间来完成计算机一秒钟之内能执行的指令。把相关的程序编码成书面指令的想法也是荒谬的：当今典型的大规模计算机程序将涉及大约一亿行计算机代码(以人类可以阅读的方式印刷出来，恐怕得好几万册书卷)。计算机可以在微秒级的时间内从内存中检索指令，而中文房间的人的运算速度恐怕要慢数十亿倍。考虑到这些实际情况，中文房间以及它所包含的一切，无法让图灵测试的询问者相信它是一个懂得中文的人：

也就是说，实际上它根本无法通过图灵测试。

还有一种对中文房间的说法是，虽然房间里的人没有表现出对中文的理解，房间本身也没有，但包含房间、人、说明书等的整个系统却有着理解。事实上，我们如果在人脑中四处挖掘，试图找到理解，也会一无所获。虽然人类的大脑某些区域负责语言理解，但在这些区域中，我们也找不到希尔勒所定义的理解力。

我对希尔勒巧妙的思维实验有着另一种解读。中文房间难题，如果从图灵测试的角度来看，是一种作弊，因为它没有把房间当作一个黑盒。只能说当我们往房间内部看的时候，在中文房间里不存在理解。而图灵测试要求我们只看输入和输出，从输入和输出来判断跟我们对话的是否真人。在我看来，争辩一个计算机系统是否“真正”理解，这是没有意义的。只要事实上它所做的事情与理解中文的人类所做的事情毫无区别，即可。

另一种反对意见是说，也许智能无法用传统的计算机实现，因为传统计算机在数学上已经被证明了具有局限性。或许你还记得，图灵的工作已经证明计算机能做什么和不能做什么——有些问题是计算机从根本上就无法解决的，但它们可以被明确界定。那么，如果人工智能所追求的智能行为在以图灵机为原型的计算机上都是无法实现的呢？图灵本人将这个观点作为强人工智能无法存在的一个可能性论据。大多数人工智能研究人员并不关心它，但是，正如我们之前经常看到的，什么是可计算问题，这一直是人工智能发展历程中的一大拦路虎。

身体还是心灵

现在我们转向意识研究中最著名的难题：身体还是心灵问题。人体和大脑中的某些物理过程会产生意识思维，但它们究竟是怎么形成的，为什么会出现呢？神经元、突触和轴突的物理世界与我们有意识的主观体验之间，到底存在什么关系？这是科学和哲学中最古老也是最艰难的问题之一。澳大利亚哲学家大卫·查尔莫斯(David Chalmers)称之为意识的难题。

有关这个课题的研究至少可以追溯到柏拉图。在《斐德罗篇》一书中，他提出一种人类行为模型，大脑中的某个控制推理的部分就像马车夫一样，控制着两匹马的缰绳——其中一匹马代表理性、高尚的渴望，

另一匹代表不合理或者无意义的欲望。一个人一生的道路取决于他的马车夫是如何控制这两匹马的。印度哲学经典《奥义书》里也出现过类似观点[147]。

把理性的自我视为马车夫是一个很好的比喻，毫无疑问，它挺可爱的——我更倾向接受控制自我内心高尚和卑劣的部分这个理论——不过它也遇到了一个常见的心灵理论上的问题。柏拉图把马车夫想象成某种“心灵”，但他所得出的结论说，人类的心灵是由另一个心灵(即马车夫)控制的，哲学上称之为侏儒问题(侏儒的意思是“小个子的人”，在这个例子中，侏儒就是马车夫)这样的解释是存在问题的，因为它实际上没有解释任何东西——只是把心灵的问题用另一个心灵问题来描述而已。

不管怎么说，“马车夫”模型是存疑的，因为它指出推理是我们行为的主要驱动力，而有大量证据表明事实并非如此。例如，在一系列著名实验中，神经科学家约翰-迪伦·海恩斯(Joln-Dylan Haynes)显然能够探测到受试者意识到自己做出最终决定的10秒前内心所做的决定[148]。

这个结果引发了各种各样的问题，但最重要的一点是，如果有意识的思考和推理并不是我们决定做什么的机制，那到底什么才是呢？

进化理论告诉我们，人体拥有的各种特征都会给我们带来进化优势。所以，按照这个理论，我们可以扪心自问，意识思维给了我们怎样的进化优势？因为根据推测，如果它没有给我们带来进化优势的话，它就不应该存在。

有一种理论认为，有意识的头脑只不过是我们身体产生各种行为的一种毫无意义的副产品，这一理论被称为副现象论。如果有意识的心灵是一种副产物，那么你的意识就不再是如柏拉图所说的，掌控缰绳的马车夫，它只是一个坐在车上的乘客，幻想自己是马车夫而已。

也有稍微中立一些的观点，认为意识并不像柏拉图说的那样在我们的行为中起主导作用，而是我们大脑进行其他活动过程中以某种方式产生的——大概是一些在低等动物大脑中不存在的过程。毕竟，据我们所知，它们并不像人类那样享受丰富的精神生活。

在接下来的内容中，我们将讨论人类意识体验的一个关键组成部分：我们的社会性，即我们理解自己和他人作为社会群体一部分的能力，以及能够思考他人和他人如何看待我们的能力。这一关键能力很可

能演变成在大型复杂社会群体中共同生活和工作的需要。为了理解这一点，我们将从英国进化心理学家罗宾·邓巴(Robin Dunbar)进行的一项著名社交大脑研究开始。

社交大脑

邓巴对一个简单的问题很感兴趣：为什么人类(以及其他灵长类生物)的大脑比其他动物都大[149]？归根结底，大脑是一个信息处理设备，它消耗了人体所产生能量的相当大一部分——通常的估算为20%。因此，灵长类生物会进化出更大的大脑，是因为它们需要处理更重要的信息。而考虑到庞大的能量需求，大脑必须产生一些实质性的进化优势。但究竟是怎样的信息处理需求，以及什么样的进化优势呢？

邓巴研究了一些灵长类生物，寻找它们可能需要增强信息处理能力的因素。例如，灵长类生物可能需要追踪环境中的食物来源，或者灵长类生物需要更大的生活范围以及觅食区域。然而，邓巴发现，和灵长类生物大脑体积关系最密切的因素是平均社会群体规模，即灵长类生物社会群体中动物的平均数量。这就表明，灵长类生物需要更发达的大脑来成功维持庞大的社交群体，更准确地说，是寻找、维持和利用群体中的社会关系。

邓巴的研究提出了一个有意思的问题：鉴于我们所知的人类大脑的平均大小，可以分析预测人类的平均群体规模是多少。通过分析，人们得出一个数值，被称为邓巴数，公认为150。也就是说，考虑到人脑的平均大小以及对其他灵长类生物的分析，我们估计人类社会群体的平均规模大约是150人，即一个人拥有稳定社交的上限人数大约是150人。邓巴数是一个能引起人好奇心的数字，随后的各种研究发现，这个数字在人类社会群体实际规模计算中反复出现。例如，新石器时代的农业村庄通常大约有150名居民。最近发现的一个有趣事实是，邓巴数可以解释我们在脸书等社交网站上积极接触的朋友数量。

邓巴数可以理解为人类大脑能够管理的人际关系的最大数量。当然，不少人互动的人数是大于这个数字的，但邓巴数是我们能够真正保持的关系数量。

简言之，如果这个分析是正确的，那么人脑的不同之处在于它是一个社会性的大脑。与其他灵长类动物相比，人脑的容量更大，因为我们

生活在庞大的社会群体中，这就要求我们有能力维持和管理大量的社会关系。

接下来的问题浮出水面了，维持和管理这些社会关系到底意味着什么？为了回答这个问题，我们将探讨一个由著名美国哲学家丹尼尔·丹尼特(Daniel Dennett)提出的观点，即我们如何用他所谓的意向立场的层面来理解和预测人们的行为。

意向立场

我们环顾四周，试图弄清楚我们看到的一切，似乎我们自然而然就能区分出智能体和其他对象。我们已经在本书中看到过智能体这个词：在之前，它指代的是构建一个人工智能程序，代表我们独立行动，理性地实现我们的偏好；而我们现在讨论的智能体，从某种意义上来说，似乎是一种跟我们有着相似属性的实体，就像有自主能力的演员。当一个孩子思考从一堆巧克力中选择哪一块，并仔细地选出时，我们看到了智能体的存在：有选择，还有有思考、有目的性、有自主性的行为。相反，当一株植物从岩石底下冒头，随着时间推移，它掀开了岩石，我们却看不到任何智能体性质：它是在以某种形式进行活动，但在活动中我们看不到思考，也看不到有意识的目的。

那么，为什么我们会把孩子挑选巧克力解释为智能体性质的行为，而把植物生长解释为一个无意识的行为呢？

为了理解这个问题的答案，想一想我们试图解释改变世界的过程时，可以从不同的层面得到不同的解释。其中之一是丹尼特所说的物理立场，它可以解释一个实体行为。在物理立场中，我们使用自然法则(物理、化学等)来预测系统的行为结果。例如，丹尼特指出，当他释放手里的一块石头时，可以使用简单的物理原理成功地预测石头会落在地上，因为石头有质量，它受到重力的作用。现在，虽然物理立场在解释这种行为的时候非常有效，但无法应用于理解或者预测人类行为，这当然是不可行的，因为人类的行为太过复杂，无法用这种方式去理解。也许原则上可行(毕竟我们最终只是一堆原子)，但应用在实践中是行不通的。就这一点而言，物理立场也不是理解计算机或者计算机程序行为的一种切实可靠的方法——典型的现代计算机操作系统的源代码长达数亿行。

另一种层面是设计立场，在这种立场下，我们根据系统应该实现的

目的进行理解，并预测系统行为，即它的设计目的。丹尼特举了一个闹钟的例子，我们不需要使用物理定律去理解闹钟的行为，我们知道它是时钟，就明白它显示的数字指的是时间，因为时钟就是用来显示时间的。同样，如果这个时钟发出刺耳的闹铃声，我们就明白它是被设定为在这个特殊时间点上开启闹钟。因为在指定时间发出刺耳的噪音也是闹钟设计的目的。这种解释不需要了解时钟的内部机制，不需要了解闹钟的具体物质构成方式、力学作用等，只需要了解它被设计出来的目的即可。

第三种层面，也是我们最感兴趣的，丹尼特称之为意向立场[150]。从这个层面看，我们将心理状态——诸如信念、欲望等归因为实体，然后使用与心理状态有关的常识去预测实体将如何行动，假设它根据自己的信念和欲望做出选择。最明显的地方在于，我们解释人类活动的时候，通常需要做出如下陈述：

> 珍妮认为天要下雨了，她希望自己能不被淋湿。
>
> 皮特想完成他的评分。

如果珍妮认为天要下雨了，又不想淋湿，我们可以预测她会穿雨衣或者带把伞，或者压根不外出。这些都是拥有上述信念和愿望的理性智能体会采取的行为。因此，意向立场是具有解释力的，它允许我们解释人们做了什么，以及他们将(可能)做什么。

请注意，与设计立场一样，意向立场对实际产生这些行为的内部构造是不关心的。这个理论同样适用于机器和人类，我们将在下面详细讨论。

丹尼特创造了意向系统这个术语，用来描述那些行为可以被有效理解和预测的实体，这些实体的行为可以归因于它们的信念、欲望和理性选择。

意向系统有着自然的层级结构，越往上越复杂。一阶意向系统有着自己的信念和欲望，但没有关于任何有关信念和欲望的信念和欲望。相比之下，二阶意向系统能够对信念和欲望产生信念和欲望。讲起来很拗口，我们举例说明：

> 一阶意向系统：珍妮相信天要下雨了。
>
> 二阶意向系统：迈克尔想要珍妮相信天要下雨了。

三阶意向系统：皮特认为迈克尔想要珍妮相信天要下雨了。

在我们日常生活中，几乎不会用到超过三阶的意向立场等级(除非我们研究人工智能的活动，比如解决一个谜题)，而且对我们大多数人而言，似乎超过五阶就很难厘清了。

社会生物的属性与我们使用意向立场密切相关，理由是它似乎能够让我们理解和预测社会中其他智能体的行为。当我们身处复杂的社会关系中，会陷入更高层次的意向思维中，于是个体的计划(不管是我们自己还是我们观察到的人)会受到其他智能体行为的影响，这里的行为可以定义为可预期的有意识的行为。因而很明显，意识思维在人类社会中普遍存在，我们也依靠它进行社交。回想一下我们在第一章中看到的爱丽丝和鲍勃的八字对话：

鲍勃："我要离开你。"

爱丽丝："她是谁？"

对于这个场景，用意向立场去解释就很简单，毫无争议，也富有说服力：爱丽丝认为鲍勃移情别恋了，并且想采取相应措施(离开她)，爱丽丝想知道那个人是谁(也许她希望挽回他)，她相信询问鲍勃会得到这个问题的答案。如果不从信念或者欲望的角度去分析，要在这种交流模式中解释清楚爱丽丝和鲍勃的角色，以及他们的思维和计划的明确特征，那会很难。

邓巴研究了人类和其他动物的大脑体积与高阶意向推理能力之间的关系。结果表明，高阶意向推理能力大致上是大脑额叶相对大小的线性函数。由于大脑的体积与社会群体的大小密切相关，因此，对于大脑自然进化的解释是为了满足在复杂社会体系中对社交推理(高阶意向推理)的需要。无论是集体狩猎、与敌对部落战斗、追求配偶(或许包括击败对手)，或是获得对盟友的影响力和领导权，理解和预测别人想法的价值不言而喻。回到邓巴的论点，更大的社会群体会对更高阶的社交推理提出更高的要求，从而解释了邓巴所确定的大脑体积和社会群体规模之间的关系。

回到我们最初的话题，意向层次似乎与意识程度密切相关。正如我们之前讨论过的，典型的一阶意向系统确实广泛存在，但更高层次的意向性就意味着更高的门槛。人类拥有高阶意向系统，但我不会接受任何人说服我相信蚯蚓也有。那么，狗呢？你也许会说，狗能够理解我的欲

望(例如，它会相信我想让它坐下来)，但如果一只狗能够进行高层次的意向性推理，那也只是一种相当有限或者经过专门训练才能获得的能力。有迹象表明，某些灵长类动物存在有限的高阶意向性推理能力。例如，长尾猴会用警告声向其他猴子表示有豹子前来袭击(威胁到了猴群的生存)。有人观察到它们也会发出欺骗性的警告声，让其他长尾猴相信自己正在被豹子攻击[151]。这种伎俩的自然解释似乎涉及更高层次的意向性推理："如果我发出警告声，其他猴子就会相信我正在被豹子攻击，然后它们就会逃跑……"当然，人们也可以提出其他解释来反驳这不代表高层次意向性推理。但是，尽管如此，这则逸事还是提供了某些非人类灵长类动物存在高阶意向性推理的可能性的证据。

以高阶意向性形式表现的社交推理似乎与意识相关。社交推理的进化足以支持复杂的社会网络和大型社会群体。但为什么社交推理需要意向性呢？我的同事彼得·米利肯(Peter Millican)提出，答案可能恰恰在于意向立场的计算效率。如果我有意识地利用自己的功能性动机——以欲望、信仰等形式，并且能够把自己想象成别人，那么这就使得我能够比其他人更有效地预测他们的行为(在实际情况和想象的情况下)。例如，如果我偷了你的食物，而我站在你的立场上去思考，可以本能地感觉到(没有经过计算)你会产生的那种愤怒，并预见你可能会采取报复措施，就会激发我去抵制偷东西的诱惑。这是一个有趣的推测，但不管人类社交推理能力和意识之间的关系如何，我们不太可能很快得到明确的答案。所以现在让我们回到我们的主题——人工智能，并考虑机器是否有能力进行社交推理。

机器有信仰和欲望吗

意向立场在人类社会中扮演着重要的角色，但它也适用于其他实体。例如，针对一个传统的开关，意向立场为我们提供了一个非常有意思的描述性解释：当开关认为我们想传输电流的时候，开关会传输电流。我们通过轻触开关来达成意愿[152]。

然而，意向立场并不是理解和预测电灯开关行为最合适的方式，在这种情况中，采用物理立场或者设计立场要简单得多。相反，对于开关的行为，是否能用意向立场进行合理的解释，这取决于它对于有电流或者没有电流的状态认知，以及我们是否有打开电灯的欲望。虽然意向立场的解释提供了关于开关电灯的准确预测，但是将它作为电灯开关与否的解释，确实不切实际。

对于是否应该用意向立场来解释机器行为，似乎存在两个主要问题：意向立场的解释是否具有合理性，以及是否具有可用性。约翰·麦卡锡是该领域颇富影响力的思想家，他对这两个问题有如下看法[153]：

> 将某些信仰、知识、自由意志、意图、意识、能力或者愿望归属于一台机器或者计算机程序是合理的，如果这种归属所表达的关于机器的信息和关于人的信息相同的话……对于已知结构的机器，比如恒温器或者计算机操作系统，智力属性的归属是最直接的。但应用于结构未知的实体，它同样有用。

这段话太难理解了，所以我们试着展开解释一下。首先，麦卡锡认为，对一台机器的意向立场的解释应该表达出关于机器的信息，就如对人的意向立场会表达出关于人的信息一样。当然，这是一个很高的要求——让人想起图灵测试的不可区分性。套用我们之前举过的例子，如果我们声称一个机器人相信天在下雨，并且机器人想保持自身干燥，那么，如果机器人表现出理性智能体应该表现的行为，就跟人一样，那么这种声称(即归属)就是合理的。所以，如果机器人能够做到的话，它会采取适当的措施避免自己淋雨受潮。如果机器人没有采取措施，我们会认为，要么它不相信在下雨，要么它不想保持自身干燥，要么它就不合理。

最后，麦卡锡指出，当我们不了解一个实体的内部构造时，意向立场是最适用于解释它的。意向立场提供了一种独立于内部结构和操作(例如，它是人、是狗还是一台机器)来解释和预测行为的方法。如果你是一个理性的智能体，有保持干爽的愿望，并且相信正在下雨，那么我可以解释和预测你的行为，而不需要了解你的任何其他信息。

通往有意识的机器之路

以上讲述的一切，跟人工智能的宏伟梦想有什么关系呢？现在，请让我冒昧地提出一些具体的建议，展望一下通往有意识的机器之路，以及我们将如何制造它。(我期待着自己年老的时候再重读这一节，看看预测是否准确。)

让我们回顾一下第五章提到的深度思维著名的雅达利游戏系统，你应该还记得，深度思维创造了一个智能体来学习大量的雅达利游戏，这些游戏在很多方面都相对简单。当然，随后深度思维的智能体进展到玩更复杂的游戏，比如星际争霸[154]。目前，这种实验值得人关注的点是：智能体能够处理具有庞大分支因子的游戏；游戏中存在有关游戏状

态或者其他玩家行为的不完美信息；游戏中执行的操作或许要等到很久以后才能得到奖励的反馈；智能体必须执行的操作不是简单的二元决策，例如打砖块，而是涉及冗长而复杂的操作序列，还可能存在与其他玩家的协作或者竞争。

这是一项令人着迷的研究，取得的进展也是令人惊叹的，但是，我们很难看到这项研究通过一些简单的进展或者突破就能制造出有意识的机器。因为这些研究正在解决的似乎不是跟意识有关的问题(当然，我不是在批评深度思维的研究，毕竟解决有意识的机器问题并不是他们研究的要点)。

鉴于以上讨论，我对如何实现这个目标提出一些初步建议。假设我们有一个机器学习程序，它可以独立学习，就像深度思维的智能体学习打砖块游戏一样。它被放在一个需要有意义的、复杂的高阶意向推理的场景中进行学习；或者是一个需要智能体说出复杂谎言的场景，这也意味着需要高阶意向推理能力。又或者是在一个场景中，智能体学会了交流，并且能够有意向地表达出自己和其他实体的意识状态。我认为，如果一个人工智能的智能体系统能够学会有意义地去做这些事情，那就是通往有意识的机器之路[155]。

我在这里想到的是莎莉-安妮测试(Sally-Anne test)，这个测试被用来帮助诊断儿童自闭症[156]。自闭症是一种严重而常见的精神疾病，在儿童时期就会表现出来[157]：

> 在儿童时期，自闭症的主要症状表现为社交和交流的发育明显不正常，典型表现是缺乏正常的灵活性、想象力和伪装本能……自闭症的社交异常主要特征表现……包括缺乏眼神交流、缺乏正常的社交意识或者适当的社交行为，即“孤独”，在互动中表现出片面性以及无法加入社交团体。

典型的莎莉-安妮测试是给被测试的孩子讲述或者表演一个小故事，通常是这样的：

莎莉和安妮同在一间屋子里，屋子里有一个篮子，一个盒子和一颗弹珠。莎莉把弹珠放在篮子里，然后离开了房间。当莎莉不在房间时，安妮从篮子里把弹珠拿出来，放进盒子里。后来，莎莉回到了房间，想玩弹珠。

然后孩子们会被问到一个问题：

“莎莉会去哪里寻找弹珠？”

比较合理的答案是“去篮子里找”，但要得出这个答案，受试者需要能够对其他人的信念做一些推理：莎莉没有看到安妮把弹珠放到盒子里，所以莎莉相信弹珠就在她放置的地方——篮子里。绝大多数自闭症儿童都会回答错误，而适龄的正常儿童几乎总能正确回答。

这个方法率先由精神病学及实验心理学家西蒙·巴伦-科恩(Simon Baron-Cohen)和他的合著者提出，以此证明自闭症儿童缺乏所谓的心智理论(Theory of Mind)的能力。心智理论能力是一种实际的、常识性的能力，成熟的成年人拥有这种能力，能够对自己和他人的精神状态(信念、欲望等)进行推理。人类并不是天生就拥有心智理论能力的，但临床上正常的人类天生就有发育它的能力。正常儿童的心智理论能力是逐步发育起来的：4岁的时候，儿童基本上能够进行推理，包括他人的态度和观点；到了青少年时期，心智理论能力就得到了充分发育。

在撰写本书之时，人们已经开始研究机器学习程序如何学习原始的心智理论能力[158]。研究人员最近开发了一个名叫ToMnet(心智理论网络)的神经网络系统，它能够学习如何对其他智能体建模，并在类似莎莉-安妮测试的情况下选择正确的行为。然而，这项研究还处于一个非常原始的阶段，解决莎莉-安妮测试问题还不足以证明人工智能拥有意识。但我认为，这是朝着正确方向迈出的第一步。它给了我们一个目标：能够通过自主学习达到人类心智理论水平的人工智能系统。

它会像我们一样吗

讨论人工智能和人类的共通之处，我们通常谈及的是大脑。这是理所当然的：大脑是人体主要的信息处理器官，当我们执行诸如解决问题、理解故事等任务时，大脑起着很重要的作用。所以我们自然会把大脑类比成无人驾驶汽车的电脑，从我们的眼睛、耳朵和其他感觉器官接收和解析感官信息，并告诉我们的手、胳膊和腿应该做什么。但这是一个极其简化的模拟过程，因为真实的大脑是一个由各种组件紧密结合在一起的系统，这个系统包含的组件极其复杂，自从生命第一次出现在这个蔚蓝色的星球上，这个系统作为一个单独的有机体已经进化了数十亿年。从进化的角度看，我们和类人猿也没什么区别——只是有意向感知的类人猿而已。所谓的人类意识，应该在这个背景下做如此理解。

我们目前拥有的能力——包括意识思维——是进化推动了原始祖先的结果[159]。正如我们的手、眼睛、耳朵的进化一样，人类的意识也在进化。人类成熟的意识并非一夜之间突现的，不像电灯那样可以突然开或者关。我们祖先的最初意识大概与蚯蚓无异，慢慢地进化成了莎士比亚。远古的祖先并没有像我们这样享受全方位的意识体验，我们也不太可能像后代那样享受到更为全方位的意识体验。进化发展的过程没有终结。

有趣的是，历史记录可以给我们一些线索，关于意识的某些元素是如何以及为何出现的。当然，有关的历史记录很少，我们不得不在大部分时间使用猜测的方式。但无论如何，这些线索很有意思。

每一个现存的类人猿——包括智人(即我们人类)——在1800万年前都有一个共同的祖先。大约在那个时候，类人猿的进化开始分支，将猩猩在约1600万年前送上了另一条进化之路，而到了600万到700万年前，智人和大猩猩、黑猩猩分道扬镳。这一次分化以后，我们的祖先开始出现区别于其他类人猿的特征：智人开始花费更多时间在地上行走，而不是在树上攀爬，最终发展到可以使用双腿直立行走。对这种进化的可能解释是气候变化减少了森林覆盖面积，迫使我们居住在树上的祖先迁往地面。离开森林的保护可能会增加被捕猎的危险，所以智人需要更庞大的社会群体来保障生存，这就要求我们进化出容量更大的大脑来支持社交推理技能，我们在之前讨论过它。

虽然早在100万年前我们的祖先就开始零星地利用火，但在大约50万年前，原始人才开始普遍使用它。火给我们的祖先带来了许多好处——它给了我们光明和温暖，吓跑了潜在的捕食者，扩大了食物的范围。然而，为了避免火灾，火的使用需要管理和维护，这就需要人们有合作的能力，以便轮流看管火堆、收集燃料等。这种合作或许促使了高阶意向推理能力(为了理解彼此的愿望和想法)的出现，还有可能催生出了语言能力。智人的语言能力似乎都是在同一时期内进化出来的，语言能力出现后，它的发展为人类进化带来的好处自然是不言而喻的。

我们无法精准地重建进化的先后顺序，以及它们带来了哪些新的能力。但普遍的研究似乎很清楚，我认为我们可以找到随着时间推移而出现的某些组成意识的部分。当然，这并不能回答意识是什么这个难题，但至少给了我们一些可能有用的线索，可以解读人类成熟的意识中一些必要的组成部分是如何以及为何出现的。它们也有可能进入死胡同，但

最终，猜想也能引领我们走向更深入的理解。比起仅仅把意识当作一个无解之谜，它们总归是提供了更多的线索。

在未来的某个时刻，我们会了解意识，就像我们现在了解驱动太阳的能量一样。在那个时候，目前关于意识的各种辩论，可能会像核物理学家正确解释太阳能量来源之前的各种理论一样有趣。

假设我们完成了我提出的假说研究，即建造出有人类心智理论能力的机器，它能够自主地学习处理复杂的高阶意向推理，能够建立和维持复杂的社会关系，能够表达自己和他人心理状态的复杂特性，那么这些机器真的会有“心智”，能够出现自我意识吗？在目前的阶段，我们无法解答这个问题。只有当我们成功建造出这样的机器以后，才能离答案更进一步。当然前提是，如果，我们有能力建造它的话。

可以想象，我们永远无法令人满意地回答这个问题，尽管我也不知道为什么会这样。在这一点上，艾伦·图灵的认知引起了我的注意，你或许还记得，图灵认为如果机器正在做的事情与“真实的人类所做的事情”无法区分，那么我们就应该停止争论机器究竟是否存在“真正”的意识。如果它能够通过我们所发明的任何合理的测试，让我们无法区分，这可能就是我们想要的结果了。

[1]　1英里≈1.61千米。

[2]　蛋白质折叠问题被列为21世纪的生物物理学的重要课题，它是分子生物学中心法则尚未解决的一个重大生物学问题。蛋白质可在短时间中从一级结构折叠至立体结构，研究者却需要花大量时间从氨基酸序列计算出蛋白质结构，而且难以得到准确的三维结构。

[3]　一部禁止使用、储存、生产和转让对人具有杀伤力的地雷，及销毁、完全禁止一切杀伤地雷的公约[160]。

后记

本书写到大约一半的某天，我和一位同事共进午餐。

“最近忙着研究什么呢？”她问我。

这是学术界的人彼此寒暄的惯例——人们总是用这种问题相互问候。我应该早点构思好一个令人印象深刻的回答。

“跟往常不太一样，我正在写一本关于人工智能的科普书。”

她哼了一声。“这世界上的人工智能科普书还不够多？你的核心观念是什么？有什么新颖的角度吗？”

这话真让人沮丧，我需要一个强有力的反击。

所以我开了个玩笑。

“这是一本关于人工智能发展史上所走过的弯路和失败方向的书。”

她看着我，脸上的笑容消失了：“那么，这本书一定很长。”

人工智能就是我的生命。从20世纪80年代中期，我还是一名学生的时候，就爱上了人工智能，时至今日，我依然狂热地爱着它。我热爱它不是因为人工智能可以给我带来巨大财富(当然那挺好的)，也不是因为我相信它能够改变世界(当然，如本书所言，我确实相信人工智能将会在诸多重要领域改变世界)。我喜欢人工智能，因为它是我所知的最迷人的学科。它借鉴了诸多学科，又神奇地推动了它们的飞速发展，包括哲学、心理学、认知科学、神经科学、逻辑学、统计学、经济学和机器人学等等。当然，从本质上讲，人工智能触及了两个人类最基本的问题：人类生存的意义，以及身为智人的我们是否独一无二。

人工智能究竟是什么，又不是什么

本书的第一个主要目标就是告诉你人工智能是什么，以及更重要的，它不是什么。这听起来很奇怪，因为可能在你看来，人工智能的本质是显而易见的。人工智能的终极梦想就是制造出拥有和人类一样的智

慧与能力的机器——拥有自我意识、自我控制、自主行为能力的机器，就如你我这般。你大概在各种科幻电影、电视和书籍中早就见过它们了。

用这种形式描述人工智能，听起来直观而明确，但当我们试图真正理解它的时候，又变得困难重重。事实上，我们根本不知道想创造的究竟是什么，或者说它需要拥有人类的哪些特质。此外，人工智能的终极目标是否是拟人化智能，长期以来也存在争议，未有定论。事实上，学术界争论得很激烈——这种人工智能是否能够存在，甚至是否可取，都没能达成共识。

基于这些原因，被称为“宏伟梦想”的拟人化人工智能很难变为现实，尽管它是诸多伟大的小说、电影或者游戏的创作基础，但它并不是人工智能研究的主流。当然，“宏伟梦想”也提出了不少相当深刻的哲学问题——我们将在本书中讨论其中一部分。除此之外，这种人工智能形式的大部分内容也仅止步于幻想，甚至有一些压根就是接近疯狂边缘的妄想——要知道，那些妄想家和卖狗皮膏药的江湖骗子之流，对人工智能的兴趣和狂热可媲美杰出的科学家。

然而，公众对人工智能的争论，以及媒体永无休止的热切关注，焦点大多在“宏伟梦想”上，即那些针对人工智能的各类危言耸听的报道和各种反乌托邦场景的描写，例如：人工智能会取代人类的工作；人工智能将比人类更聪明，然后脱离人类的控制；超级人工智能将会在程序出错的时候消灭人类……不少发表在大众媒体上的有关人工智能的文章严重缺乏相关知识，甚至有些还与人工智能毫不相干。从技术角度来讲，那些文章大多是垃圾，不管它们读起来多么吸引人。

我希望通过本书来改变这些观点。我要告诉人们，人工智能究竟是什么，研究人工智能的科研人员究竟在做什么，以及他们是如何做到的。在可预见的未来，人工智能的技术实现与所谓的“宏伟梦想”大相径庭。它也许不是那么引人注目，但是——正如我在本书中将要展示的那样——其本身就是一门令人兴奋的学科。当今人工智能研究的主流是让机器完成一些目前必须借助人脑(当然，有可能还有人类的肢体)完成的特定的任务，还有一些传统计算机技术无法解决的任务。21世纪以来，人们在这一领域取得了重大突破，这也是人工智能成为热门话题的原因。举个例子，20年前，自动翻译工具还只存在于科幻小说中，但在过去的10年里，它已经成为人们可以实实在在运用的日常工具。虽然自动

翻译工具有很多局限，但每天全球仍有数百万人在使用它。在未来的10年内，我们将用到质量更高的、能够实现同声传译的语言工具。而增强现实(AR)技术将改变我们感知、理解和连接世界的方式。无人驾驶汽车是一个富有现实意义的人工智能应用。医疗和保健领域的人工智能应用则会使所有人受益匪浅：在识别诸如X光和超声波图像等领域，目前已证实人工智能在判断肿瘤等异常现象方面优于人类；加载了人工智能技术的可穿戴设备，能够长期监测人体健康状况，可以给出使用者罹患心脏病、高血压甚至阿尔茨海默病的预警。这才是人工智能研究人员真正致力于解决的问题，这才是人工智能让人兴奋不已的地方，这才是人工智能的故事应该讲述的东西。

要理解如今的人工智能是什么，以及为什么基本上人工智能与所谓的“宏伟梦想”没多大关系，我们还需要理解为什么制造人工智能如此困难。在过去的60年里，大量的研究工作(还有海量的资金)流向人工智能研究领域，然而，遗憾的是，梦想中的全能机器人管家还不太可能在短时间内出现。那么，为什么人工智能研发如此困难？为了弄清楚这个问题，我们需要从最基础的层面开始了解计算机是什么，计算机能做什么。这就把我们带向了数学中最深奥的领域之一，以及20世纪最伟大的科学家之一艾伦·图灵的成就。

人工智能的历史

本书的第二个主要目标是告诉大家人工智能从初创开始的故事。既然讲故事，那就得有情节，据我所知，所有现存的故事实际上只有七种基本情节[1]。那么，哪一种模式最适合讲述人工智能的故事呢？我的同事非常希望这是一个“白手起家”的过程，对于少数聪明(或者说幸运)的人而言，确实如此。但出于稍后能够讲述的原因，我们可以更合理地将人工智能的发展故事视为“杀魔除怪”——在这里，“魔怪”是一种抽象的数学理论，被称为“计算复杂性”，它才是人工智能的问题如此难以解决的真正原因。“探寻旅程”也不错，因为人工智能的故事很像中世纪的骑士寻找圣杯的过程：充斥着宗教式的狂热、无可救药的乐观主义、错误百出的线索、钻不出的死胡同和痛苦的失望。其实，归根结底，描述人工智能故事情节的最佳模式应该是“起起落落”，因为就在20年前，人工智能还是一个特别小众的领域，其学术声望遭到诸多质疑，而后则迅速上升为当代科学最有活力、最炙手可热的领域。不过，人工智能故事的情节模式更准确地说应该是“起落起落起”。半个多世纪以来，人们对人

工智能的研究从未间断，在此期间，研究人员一而再、再而三地发表声明，宣称做出了重大突破，智能机器的宏伟梦想触手可及。结果呢，这一切全都被证明是无可救药的盲目乐观。人工智能就在虚假繁荣和过度萧条之间起起落落，甚至因而臭名昭著——在过去40年中，至少出现过三次这样的循环。延伸到过去60年里，人工智能领域受到了好几次严重冲击，几乎将它就此毁灭。然而，每一次毁灭以后，它都坚强地再起。如果你认为科学就是从无知到开悟的有序发展，那你应该感到震惊。

现在，人工智能又一次进入了繁荣阶段，人们对人工智能的热情达到了极点。在这个充满期待的时刻，我认为，至关重要的是，那些令人生厌的故事该被一次又一次提起。人工智能的研究人员不止一次地认为，他们发现了推动人工智能走向某种宿命的神奇要素。对当前人工智能引起的令人屏住呼吸、睁大双眼的热切讨论，同样也是这个调调。据报道，谷歌公司CEO桑达尔·皮查伊(Sundar Pinchai)曾宣称，“人工智能是人类迄今为止最重要的事情之一，甚至比火、电还要有意义1”。这句话继承了吴恩达此前提出的主张，即人工智能是“新的电能”2。不过重点在于，我们得铭记历史中的傲慢。没错，真正的进步是有的；没错，它们是值得庆贺并激动人心的——但它们并没有带领我们走向人工智能的终极目标：有自我意识的机器。作为一个有着30年工作经验的人工智能研究人员，我主张对该领域的研究保持高度谨慎的态度，尤其是对所宣称的重大突破，我会持审慎和怀疑的态度。人工智能研究现在最迫切需要的，是谦卑。这让我想起了古罗马凯旋的故事——凯旋的将军身后有一名奴隶跟随，奴隶手持金冠置于他头上，同时在他耳边低语一句拉丁文“memento homo”(谨记，你只是一个凡人)。

因此，本书的第二部分按大致的时间顺序讲述了人工智能发展的历史，毫不避讳，毫不遮掩。人工智能的故事开始于第二次世界大战之后第一台计算机的诞生，我将带你经历所有人工智能的繁盛周期，从“黄金年代”开始，那是一段肆无忌惮的乐观年代，那段时间似乎所有的战线都取得了飞快的进步。接下来是“知识时代”，那时候的构想是让机器获取我们人类所拥有的一切知识。然后是“行为时代”，那时候人们坚决主张机器人是人工智能的核心。再往后是现在的“深度学习时代”。每一个时期都有其独特的构想，还有那些贡献非凡想法的人。

人工智能的未来

虽然我得说，在当下对人工智能狂热追捧的氛围中，我们应该冷静

下来，认识到人工智能发展史其实充斥着各种失败，走过许多弯路(盲目乐观和过度兴奋应该降降温)，但我同样相信，我们有实实在在的理由来对人工智能的未来保持乐观。在科学上有了名副其实的重大突破，再加上大数据和变得廉价的强大计算力，通过十来年的发展，人工智能系统已经初具雏形——这一领域的创始人将之誉为奇迹。所以，我的第三个主要目标是告诉大家人工智能系统现在能做什么，以及在不久的将来它能做什么——另外，指出它的局限性。

这会引领我们继续探讨人工智能给人类带来的恐惧，就像之前我提到的那些，主要是反乌托邦场景里的一些东西，比如人工智能会占领世界之类的。人工智能的最新进展确实引起了一些值得我们关注的问题，比如人工智能时代就业形式的改变，人工智能技术如何影响人权——关于机器人代替人类统治现实世界(或者其他现实领域)的讨论，恰好把这些非常重要、非常真实的问题挤出了新闻头条。所以，再强调一次，我希望改变这种论调。我希望引导大家关注真正值得关注的领域，并尽可能清晰地指出我们究竟应该害怕什么，又没必要担心什么。

最后，我希望大家可以轻松一些。因此，在本书最后的章节，我们回归到了拥有自我意识、自我控制、自主行为能力的机器的“宏伟梦想”话题上。我们详细探讨了这个梦想，并了解实现这一梦想意味着什么，这样的机器会以怎样的状态实现——它跟我们会不会高度相似。

这本书能给你带来什么

本书围绕了以上提到的几个主要目标展开：告诉你什么是人工智能，为什么它实现起来非常困难；告诉你人工智能的故事——在每个繁荣时期推动它发展的想法和杰出科学家们；最后，展示人工智能现在能做什么，不能做什么，并谈一谈人工智能的远景——通往有自我意识的机器之路。

写这本书的乐趣之一是摆脱了通常很重要但是令人厌烦的学术写作惯例。因此，我在本书中只标出了一些重点引用，参考文献相对较少。

我不光摒弃了大量的参考文献，还避免了诸多技术细节——我的意思是那些数学细节。

我希望，在读完本书后，你会对推动人工智能发展的主要思想和概念有一个宽泛的了解。事实上，这些想法和概念大多都是高度数学化

的。但我高度赞同斯蒂芬·霍金的名言：科普书里的每一个方程式都会让它流失一半的读者。当然，对那些自觉能够接受高难度挑战的读者，我准备了一些较为深入探讨技术思想的附录，以及可供深入阅读的思路。

本书算是精挑细选之作，我真的已经尽力了：人工智能是一个庞大的领域，要绝对公正地对待过去60年来影响人工智能的各种不同思想、流派和传统，那完全不可能。我只能试图寻找出自己认为构成人工智能宏伟蓝图的每一根主要的经纬线，让它们交织成人工智能的故事。

最后，我要提醒大家，这不是一本教科书。阅读本书不会让你有能力创办一家人工智能公司，也不会让你具备成为谷歌或者脸书技术员的技能。你能从本书中得到的，是理解人工智能的概念，以及它未来的走向。我希望你在阅读之后，能正确了解到人工智能的真实现状，并且，能够帮助我改变大众对人工智能的论调。

2019年5月写于牛津大学

[1] 《七种基本情节：我们为什么讲故事》是荣格学说研究者克里斯托弗·布克(Christopher Booker)于2004年出版的书籍，书中研究了各类故事和传说的模式以及其心理学含义。他将七种基本情节定义为：杀魔除怪、白手起家、探寻旅程、远行与回归、喜剧、悲剧、重生。

致谢

在此感谢我的代理人费利西蒂·布莱恩(Felicity Bryan)和编辑劳拉·斯蒂克尼(Laura Stickney)在项目研究过程中给予我的支持、鼓励和建议。另外，我得补充一句，他们真的很善良，还有非凡的耐心。

特别鸣谢：伊恩·古德费洛(Ian J. Goodfellow)、乔纳森·什伦斯(Jonathon Shlens)和克里斯蒂安·塞吉(Christian Szegedy)，允许我在本书中引用他们论文《解释和利用对抗性范例》中的“熊猫/长臂猿”图片(图17)；彼得·米利肯允许我改变我俩共同撰写的论文中的一些素材；苏巴拉奥·坎班帕提(Subbarao Kambhampati)为我的一些重要论点提供了有用的反馈，并且向我指出了思考帕斯卡赌注[1]的思路；奈杰尔·沙博尔特(Nigel Shadbolt)与我讨论，给我鼓励，并提供了不少有关人工智能的奇闻异事；里德·史密斯(Reid G. Smith)提供了MYCIN系统的相关教学素材，并且允许我改编。

感谢我的同事、学生，还有那些运气不好被我逼着短时间内阅读完本书原稿的朋友，他们都给了我非常有价值的反馈。感谢阿尼·卡林斯库(Ani Calinescu)、蒂姆·克莱门特-琼斯(Tim Clement-Jones)、卡尔·本尼迪克·弗雷(Carl Benedikt Frey)、保罗·哈伦斯坦(Paul Harrenstein)、安德鲁·霍奇斯(Andrew Hodges)、马蒂亚斯·霍尔韦(Matthias Holweg)、威尔·赫顿(Will Hutton)、格雷厄姆·梅(Graham May)、艾达·梅霍尼克(Aida Mehonic)、彼得·米利肯、史蒂夫·纽奥(Steve New)、安德烈·尼尔森(André Nilsen)、詹姆斯·保林(James Paulin)、艾玛·史密斯(Emma Smith)、托马斯·斯泰尔斯(Thomas Steeples)、安德烈·斯特恩(André Stern)、约翰·桑希尔(John Thornhill)和基里·沃尔登(Kiri Walden)。当然，本书难免还存在诸多谬误，那完全是我的责任。

我很荣幸收到欧盟科学研究协会卓越贡献奖291 528欧元的奖金，我和我的团队依靠这笔奖金完成了2012—2018年的研究项目。我很庆幸拥有由朱利安·古铁雷斯(Julian Gutierrez)和保罗·哈伦斯坦领导的才华横溢、精力旺盛、亲密无间的研究团队，他们使我的工作变得无比高效和愉悦。非常期待我们能在未来的日子里继续合作。

[1] 帕斯卡赌注(Pascal's wager)是法国数学家、物理学家、思想家布莱士·帕斯卡

(Blaise Pascal)在其著作《思想录》中表达的一种论述，即：我不知道上帝是否存在，如果他不存在，作为无神论者我得不到任何好处；但是如果他存在，作为无神论者我将有很大的坏处。所以，我宁愿相信上帝存在。

词汇表

A*：人工智能中使用最广泛的启发式搜索方法，是斯坦福研究所SHAKEY项目的一部分，于20世纪70年代早期开发。在A*之前，启发式搜索是一种相当特殊的技术。A*让启发式搜索建立在坚实的数学基础上。

AlexNet：一个突破性的图像识别系统。该系统在2012年实现了图像识别的巨大改进，是一个具有里程碑意义的深度学习系统。

AlphaGo：一个突破性的围棋游戏系统，由深度思维的团队开发。2016年3月，在韩国首尔举行的一场比赛中，AlphaGo以4：1的成绩击败世界围棋冠军李世石。

ceteris paribus preferences：即“尽可能保持其他条件不变”，这个概念是指我们在给人工智能系统说明自己的偏好时，希望它执行任务的前提是“尽量保持其他事物不变”(即尽可能让其他的事物保持原来的状态)。

Cyc工程：基于知识的人工智能时代中，最著名也是得到负面评价最多的实验。它试图通过提供一个受过合理教育的人类所拥有的关于世界的一切知识来构建一个通用人工智能系统，但最终失败了。

Cyc假说：一种假设，即人工智能主要的问题是知识问题。一个拥有全面知识库的人工智能系统可以等同于通用人工智能。

DENDRAL：一个经典的早期专家系统，帮助用户识别未知的有机化合物。

ELIZA：20世纪60年代由约瑟夫·魏岑鲍姆开发的会话式人工智能开创性实验，ELIZA用简单的封装脚本来模拟心理医生。

HOMER：20世纪80年代开发的一种智能体，它在模拟的“海洋世界”环境中运行。HOMER可以使用(一部分)英语与人交互，被指派在海洋世界完成任务，并对自己的行为有一些常识性的理解。

ImageNet：一个由李飞飞开发的带有标签的大型在线图像档案库，在训练深度学习程序为图像添加标签的领域有着深远影响。

LISP：在符号人工智能时代被广泛应用的编程语言，由约翰·麦卡锡开发。LISP机器是专门设计来运行LISP编程语言的计算机。另请参见PROLOG。

MYCIN：20世纪70年代出现的经典专家系统，它充当医生的助手，诊断人类血液疾病。

NP完全问题：一类无法有效解决的计算问题。NP完全性理论是20世纪70年代发展起来的，在此期间许多人工智能问题被发现是NP完全问题。另请参见P与NP问题。

PROLOG：一种基于一阶逻辑的编程语言，在基于逻辑的人工智能时代特别流行。

P与**NP**问题：P代表“多项式时间”，NP代表“非确定性多项式时间”，从技术上讲，P与NP问题指的是在非确定性多项式时间内可以解决的问题是否可以在多项式时间内解决。NP完全问题是否一定能够在多项式时间内解决？这是当今数学中最大的难题之一，在短期内可能无法解决。

R1/XCON：20世纪70年代美国数字设备公司开发的经典专家系统，旨在帮助他们配置VAX系列计算机。这是早期人工智能的盈利案例。

SHAKEY：20世纪60年代末，斯坦福研究院在自主机器人方面进行的有重大意义的实验，开创了几项关键的人工智能技术。

SHRDLU：人工智能黄金时代出现的著名系统，后来因为专注于模拟的微观世界而广受非议。

STANLEY：赢得2005年美国国防高级研究计划局顶级挑战赛冠军的无人驾驶汽车，由斯坦福大学开发，自动驾驶大约140英里，平均时速19英里。

STRIPS：一个影响深远的计划系统，在斯坦福研究院的人工智能项目中使用。

阿西洛玛人工智能准则：2015年和2017年，人工智能科学家和评论员在加利福尼亚州阿西洛玛举行了两次会议，制定了一套关于人工智能伦理的准则。

包容式体系结构：一种行为人工智能时代的机器人体系结构，将机器人可能的行为组织成一个层次结构，越是底层的行为，优先级就越高。

贝叶斯定理/贝叶斯推断：贝叶斯定理是概率论的一个核心概念，在人工智能中，它为我们提供了一种在得到新数据或者证据时调整人工智能信念库的方法。最关键的是，新的数据可能只是“噪音”或者不够确定的数据。贝叶斯定理为我们提供了处理此类不确定信息的正确方法。

贝叶斯网络：一种知识表述方案，用于捕获基于概率的数据之间产生的复杂网络。使用基于贝叶斯定理的贝叶斯推断来构成。

本体工程：在一个专家系统中(更普遍地说是在一个基于知识的人工智能系统中)，这是一个定义概念词汇表的任务，这些词汇表被用来表示系统中的知识。

博弈论：一种基于全局的推理理论，在人工智能中被广泛用于构造人工智能系统和其他系统交互的框架。

不可判定问题：从精确的数学意义上来讲，某个问题无法用计算机(或者更确切地说，图灵机)来解决，即为不可判定问题。

不确定性：人工智能普遍存在的一个问题。我们接收到的信息很少是确定的(即确切为真或者为假)，通常会具有不确定性。同样，当我们做决策的时候，也很少确切地知道决策后果是什么：通常有多种可能性结果，只是每种结果出现的概率不同。因此，处理不确定性是人工智能的一个基本主题。另请参见贝叶斯定理及附录C。

不确定性选择：在这种情况下，我们所做的选择可能有多种结果，而我们所知的，是对于每一个可能的选择，其结果发生的概率。参见预期效用。

不通情理的实例化：人工智能系统按照使用者的要求达成任务，但并没有按照使用者预期的方式去做。

(神经网络的)不透明性：神经网络的专业知识都被编码在一系列的权重数字中，我们无法分辨这些权重的“含义”。这意味着目前的神经网络不能解释或者证明它们的决策。

常识推理：这是一个宽泛的术语，基本上人类都有能力做出有关世界的常识性的推断，但这对基于逻辑的人工智能来说非常困难。

城市挑战赛：2007年美国国防高级研究计划局顶级挑战赛的后续挑战赛，在城市挑战赛中，无人驾驶汽车必须能够自主穿越城市环境。

触发：在基于知识的人工智能系统中，如果在工作存储器中的信息与规则的前因条件正确匹配，那么规则就会被触发，从而允许我们将规则的后果结论添加到工作存储器中。

触发阈值：即人工神经元被触发的临界值。人工神经元接收大量的信号输入，其中可能只有一部分输入被激活，当被激活的输入的权重值超过触发阈值时，神经元会被“触发”(即产生一个输出)。

初始状态：在解决问题中，初始状态描述了我们在执行任务之前呈现的状态。另请参见目标状态。

传感器：给机器人提供原始感知数据的装置。典型的传感器有摄像机、激光雷达、超声波测距仪和碰撞探测器。解释原始感知数据是人工智能的重大挑战之一。

道德机器：一个在线实验，在这个实验中，用户被问及在各种电车难题中应该做出什么样的选择。

道德智能体：如果一个实体能够理解其行为的结果以及有是非的观念，并且能够因此对其行为负责，那它就是一个道德智能体。主流观点认为，人工智能系统不应该，事实上也不可能被视为道德智能体。道德责任在于构建和运行人工智能系统的人，而不是系统本身。

电车难题：伦理学中的一个问题，最早出现于20世纪60年代：如果你无作为，有5个人会死亡，而如果你采取行动，那么另外1个人会死亡。你应该行动吗？通常在无人驾驶汽车背景下讨论，尽管大部分人工智能研究员认为无关紧要。

顶级挑战赛：美国国防高级研究计划局组织的无人驾驶汽车竞赛，

2005年10月，由STANLEY机器人获得胜利，标志着无人驾驶汽车时代的到来。

对抗性机器学习：机器学习的一个分支，人们给程序输入容易导致错误输出的数据，试图欺骗机器学习程序，虽然这些输入的结果对人类而言是“显而易见”的。

多层感知器：多层神经网络的早期形式。

多层神经网络：将神经网络组织成一系列多层级网络的标准方法，每层的输出作为下一层的输入。早期神经网络研究的一个关键问题是没有办法训练多层神经网络，而单层神经网络的能力又十分有限。

多智能体系统：多个智能体相互作用的系统。

反向传播：训练神经网络的一种重要算法。

反向推理：在基于知识的人工智能系统中，我们从试图建立的目标(例如“该动物是食肉动物”)开始，试图通过使用我们所掌握的数据来验证目标是否合理(例如“如果动物会吃肉，那么该动物是食肉动物”)。与正向推理相对。

分支因子：在解决问题的时候，每次做决定时必须考虑的备选方案的数量。在玩游戏的时候，分支因子即你在任何棋盘位置上平均可能移动的次数。在井字棋游戏中，分支因子约为4；在国际象棋中，约为35；在围棋中，约为250。较大的分支因子会导致搜索树很快变得庞大到难以想象，因此需要使用启发式搜索来优化。

符号人工智能：一种人工智能类型，包括明确的推理建模和规划过程。

副现象论：一种意识研究的观念，即意识是身体内部机械运动的结果，是物质运动之后伴随出现的次生现象。

感受性：个人心理体验。比如闻到咖啡味，或者在热天喝冷饮之类。

感知：了解周围环境的过程。它是符号人工智能的根本症结所在。

感知器/感知器模型：一种神经网络，起源于20世纪60年代，但至今仍是有价值的。对感知器的研究在20世纪70年代早期就逐渐停止，因为当时有研究表明单层感知器模型所能学习的东西有着严重局限性。

高级编程语言：一种隐藏程序运行时实际计算机底层操作的编程语言。高级编程语言至少在原则上是独立于机器的：同一个程序可以在不同类型的计算机上运行，比如Python和Java。约翰·麦卡锡开发的LISP就是早期的高级编程语言例子。

功利主义：应该选择让社会福利最大化的行动的观点。在电车难题中，功利主义者会选择杀死一个人来挽救五个人的生命。另请参见美德伦理学。

工作存储器：在专家系统中，工作存储器里包含了正在解决问题的信息(与编码在规则中的知识相对)。

规则：一种离散的知识。以“如果……那么……”的形式表示，例如，规则“如果动物有乳房，那么该动物为哺乳动物。”它告诉我们，如果我们知道某动物是有乳房的，那么我们可以据此得到新的信息，即该动物为哺乳动物。

后果结论：专家系统的“如果……那么……”规则中，“那么”后面跟随的部分即为后果结论。例如，在规则“如果动物有乳房，那么该动物是哺乳动物”中，后果结论是“该动物是哺乳动物”。

极大极小值搜索：游戏开发使用的关键搜索技术之一，在游戏中，你试图使自己的利益最大化，而你的对手会试图使你的利益最小化。另请参见搜索树。

积木世界：一个模拟的“微观世界”，人工智能在其中的任务是摆放各式各样的物体，如方块或者盒子。其中最著名的是SHRDLU所在的积木世界，随后因为抽象了人工智能系统在现实世界中所面临的诸多真正困难的问题——尤其是感知问题——而广受批评。

机器人三定律：科幻作家艾萨克·阿西莫夫在20世纪30年代提出的三条定律，是一种规范人工智能行为的伦理框架。虽然它们非常巧妙，但实际上不可能直接在人工智能编码中实现，甚至无法被明确定义。

机器学习：智能系统的核心功能之一。机器学习程序学习输入和输出之间的关联，而不需要明确告诉它按照怎样的步骤去执行。神经网络和深度学习是机器学习的常用方法。

基于逻辑的人工智能：一种人工智能系统构建方法，其中的智能决策被简化为逻辑推理，例如一阶逻辑。

基于行为的人工智能：1985年至1995年期间，一种获得了广泛关注的取代符号人工智能的系统。该系统在构建的时候更为关注系统应该表现出的行为，再去考虑这些行为是如何关联的。包容式架构是构建基于行为的人工智能系统最流行的方法。

基于知识的人工智能：1975年至1985年间，人工智能的主流范式。使用明确的知识来解决问题，通常以规则的方式编码。

基于智能体的交互界面：通过人工智能驱动的软件智能体来实现计算机界面操作的想法。软件智能体与使用者共同工作，智能体是一个主动的助手，而不是像常规计算机应用程序那样只能被动地等待使用者告诉它做什么工作。

监督式学习：机器学习最简单的形式，通过向程序展示输入和期望输出的实例来训练程序。另请参见训练。

奖励：强化学习程序对其行动的反馈，奖励可能是正面的，也可能是负面的。

脚本：20世纪70年代发展起来的一种知识表述方案，旨在捕捉常见情况下常规事件的序列。

精神状态：意识的一个重要组成部分，包括欲望、信念和偏好等。另请参见意向立场。

可判定问题：即可以使用算法解决的问题。

莱特希尔报告：20世纪70年代初，英国的一份关于人工智能的报告，对当时的人工智能研究提出了严重的质疑。该报告直接导致了人工智能研究经费大幅削减，普遍认为该报告是导致人工智能寒冬出现的因素之一。

逻辑：推理使用的正规框架。另请参见一阶逻辑和基于逻辑的人工智能。

逻辑编程：一种编程方法，我们简单地陈述对问题的了解和我们的目标，然后剩下的由机器去执行。另请参见PROLOG。

旅行机：20世纪90年代中期出现的一种典型的智能体设计。它将智能体的控制系统分为三层，分别负责反应、计划和建模。

美德伦理学：当面临伦理问题时，确定一个道德智能体，他体现了我们所注重的道德原则，然后他的选择就是我们应该去做的选择。

目标状态：在问题解决中，目标状态描述了我们成功完成任务时希望看到的状态。

纳什均衡：博弈论里的一个核心概念，如果任意一位参与者在其他所有参与者的策略确定的情况下，其选择的策略都是最优的，那么这个组合就被定义为纳什均衡。

逆向强化学习：机器学习程序观察人类的行为，试图从观察中学习到奖励系统。

判定问题：判定问题是指答案可以用是/否来回答的数学问题。例如“16的平方根是4吗？”和“7920是质数吗？”之类。艾伦·图灵解决了判定问题，即是否所有的数学判定问题都是可判定的。他指出有一些判定问题(尤其是停机问题)无法用算法来解决，这种类型的问题即是不可判定问题。

偏好/偏好关系：对于所有可供选择的事物，你按照偏好程度进行排序。如果想要一个智能体代表你行事，它需要知道你的偏好，才能为你做出最佳选择。

普罗米修斯工程：20世纪80年代欧洲无人驾驶汽车技术的开创性实验。

奇点：机器智能超越人类智能的假设节点。

启发式搜索：一种利用“经验法则”来决定集中搜索方向的方法，启发式搜索是基于经验法则的，所以无法保证搜索方向一定正确。另请参

见A*。

嵌入(智能体)：行为人工智能时代的一个中心思想，大意是指人工智能的进步需要开发真正嵌入某个环境中并且能够作用于该环境中的智能体，而不是脱离实体开发(专家系统通常就是这样)。

前提：在逻辑上，前提是你最初掌握的知识，你要使用逻辑推理等方式从前提中得出结论。

前因条件：专家系统的“如果……那么……”规则中，“如果”后面跟随的部分即为前因条件。例如，在规则“如果动物有乳房，那么该动物是哺乳动物”中，前因条件是“动物有乳房”。

强化学习：一种机器学习的形式，让智能体在环境中进行活动，并以奖励的形式接受对其行为的反馈。

强人工智能：强人工智能的目标是建立一个真正有思维和意识等的人工智能系统，另请参见弱人工智能和通用人工智能。没有人知道强人工智能是否能实现，以及它会是什么样子。

(神经网络中的)权重：在神经网络中，神经元之间的连接(轴突)被赋予了权重数字。权重越大，连接对神经元的影响就越大。神经网络最终归结于这些权重值，而训练神经网络的目的就是找到合适的权重值。

人工智能的黄金年代：人工智能研究的早期，大约从1956年到1975年(紧接着就是人工智能寒冬)。这一时期的主要研究集中在“分而治之”的方法上：构造能够展示智能行为组成部分的系统，希望它们在以后能够被集成为通用人工智能。

人工智能寒冬：指20世纪70年代早期，莱特希尔报告发表后的一段时间。该报告对人工智能极尽批判之事。在人工智能寒冬中，研究经费大幅削减，整个人工智能领域都遭受质疑。人工智能寒冬出现在基于知识的人工智能系统兴盛之后。

软件智能体：在软件环境中而不是像机器人那样在物理环境中存在的智能体。可以把软件智能体视为软件机器人。

弱人工智能：构造虽然没有具备人类的理解力(意识、思维、自我意识等)，但是可以模拟出相应能力的程序，即为弱人工智能。另请参见强

人工智能和通用人工智能。

莎莉-安妮测试：一种心理测试，目的是确定被测试者是否有心智理论能力，即对他人的欲望和信念进行推理的能力。这项研究最初是为了测试自闭症。

社会福利：“社会的总体幸福感”，社会整体状况如何，是社会福利的一种衡量方式。

设计立场：试图根据某个实体的设计意图来理解和预测它的行为。例如，时钟是用来显示时间的，所以我们可以把它显示的数字理解为时间。与物理立场和意向立场形成对比。

实体程序：人工智能程序中表示偏好的标准技术。给所有可能出现的结果附加数值表示偏好程度，称为实体程序。然后，人工智能系统试图计算行动过程，得出效用最大化的结果，即偏好程度最高的结果。另请参见预期效用和预期效用最大化。

深度学习：21世纪出现的推动机器学习研究兴起的突破性技术。其特点是使用更深层次结构、互联性更高的神经网络，以及更庞大、更精心策划的训练数据和一些新技术。

深度优先搜索：一种用于解决问题的搜索技术，在深度优先搜索中，并非逐层展开搜索树，而是沿着搜索树的一个分支向深度展开。

神经网络：一种利用“人工神经元”进行机器学习的网络构造，深度学习中的基本技术。另请参见感知器。

神经元：即神经细胞。神经元彼此相连，并通过轴突传导信息。大脑的基本信息处理单元，也是神经网络的灵感来源。

身体还是心灵问题：科学中最基本的问题之一，即大脑/身体中的物理过程如何与意识体验相关联？

搜索/搜索树：一种基本的人工智能问题解决技术，计算机程序试图从初始状态开始，采用有限的步骤，寻找如何转化成目标状态的方法。

算法偏见：人工智能系统在做决策的时候可能会表现出偏见。其原

因可能是使用了有偏见的数据集进行训练，或者是软件设计不当。偏见可能以分配伤害或代表性伤害的形式出现。

特征：分段的数据组成，机器学习程序依赖它进行学习以做决策。

特征提取：在机器学习中，决定在一个数据集中选择哪些属性作为训练数据。

梯度下降：训练神经网络时使用到的一种技术。

通用人工智能：这是个宏伟的目标，即建立一个拥有人类所拥有的全部智能能力的人工智能系统：计划、推理、参与自然语言对话、开玩笑、讲故事、理解故事、玩游戏——所有的一切。

通用图灵机：现代计算机的原型。虽然图灵机通常只能编码一种特定的步骤或者算法，但通用图灵机可以执行任意的算法。

突触：神经元之间的“连接点”，让它们能够交流信息。

图灵测试：由艾伦·图灵提出的一种测试，用来解决机器是否能“思考”的问题。如果你与一个实体互动一段时间后，你不能分辨出它是一台机器还是一个真人，你就应该接受它具有类似人类智能的结论。它很有创意，也很有影响力，但是在人工智能测试领域，不必把它奉若圭臬。

图灵机：一种包含着解决特定数学问题的特殊步骤的机器。任何可以用计算机解决的问题都可以用图灵机解决。由艾伦·图灵发明，用来解决判定问题。另请参见通用图灵机。

(不完备的)推理：在逻辑学中，得出的结论并不是前提所能保证得出的。另请参见(完备的)推理。

(完备的)推理：如果从前提中能够正确得出结论，那么推理就是完备的。

推理机：专家系统中负责推理的部分，从工作存储器的规则和事实中获得新知识。

维度诅咒：在机器学习中，包含更多特征的训练数据会导致训练数量剧增。

危害评估风险工具：一个机器学习系统，用来帮助英国杜伦的警察决定是否应该拘留某人。

威诺格拉德模式：图灵测试中的一种模式。两个句子只有一个单词不同，但却具有完全不同的意思。这项测试需要被测试者理解不同之处。威诺格拉德模式中，必须理解文本以后才能正确回答，旨在识别出在图灵测试中使用一些小伎俩来欺骗询问者的被测试者。

问题解决：在人工智能中，问题解决意味着寻找到正确的行动顺序，将问题从初始状态转换为目标状态。搜索是人工智能中问题解决的标准方法。

物理立场：试图通过物理结构和物理定律来预测和解释一个实体的行为。与设计立场和意向立场形成对比。

乌托邦主义者：相信人工智能和其他新技术将带领我们走向一个乌托邦式的未来的人(技术将使人们从工作中解脱等)。

狭义人工智能：和通用人工智能相对，构建专注于解决非常具体问题的人工智能系统，而不是试图让人工智能系统拥有人类全部的智能能力。这个词主要见诸媒体，人工智能研究界极少使用。

先验概率：在得到进一步信息之前，你先假设一个概率。在这个意义上，“先验”意味着“在得到更多信息之前”。

信念：人工智能所掌握的有关其环境的信息。在基于逻辑的人工智能中，系统所拥有的信念即知识库和工作存储器中的信息。

信用分配：机器学习中出现的一个问题：决定机器学习程序的哪些动作是好的，哪些是坏的。例如，你的机器学习程序输掉了一盘棋，它如何得知具体哪一步是导致输棋的关键步骤呢？

心智理论：临床上表现正常的成年人对他人的精神状态(信念、欲望、意图等)进行推理的普遍能力。另请参见意向立场和莎莉-安妮测试。

(机器学习中的)训练：机器学习程序的任务是学习输入和输出之间的关联，而无须被告知它们之间是如何关联的。为了做到这一点，程序通常是通过给出输入和相应输出的示例来进行训练。另请参见监督式学

习。

演绎：即逻辑推理，从现有的知识中获取新的知识。

易处理：如果某个问题能够被一种有效算法所解决，这个问题可以被认为是易处理的。NP完全问题就不是易处理问题：我们没有有效算法解决它们。另请参见P与NP问题。

一阶逻辑：一种通用的语言和推理系统，为数学推理提供了精确的基础。在基于逻辑的人工智能范式中得到广泛研究。

意识的难题：理解物理过程如何以及为何会导致主观意识体验的问题。另请参见感受性。

意向立场：预测和解释某个实体的思想，将其归因到实体的心理状态上，如信念或者欲望，并假定该实体根据这些信念和欲望理性地行动。与物理立场和设计立场形成对比。

意向系统：任何可以使用意向立场来描述的系统。

涌现性质：组成系统的每个组件都会表现出某种特性，当这些特性发生交互的时候，通常会以一种意外的或者不可预测的方式产生新的特性。

预期效用：在不确定条件下的判定问题中，某一特定决策方案的预期效用是指在此选择下平均获得的收益。

预期效用最大化：在不确定条件下的决策中，当在多个方案中做出选择时，理性的智能体将选择平均收益最大的方案，即预期效用最大化方案。

语义网：一种知识表述方案，使用图形表示法来捕获概念和实体之间的关系。

正向推理：在基于知识的人工智能系统中，从信息中推理得出结论。与反向推理相对。

制订计划：即寻找到将初始状态转变成目标状态的一系列操作。另请参见搜索。

知识表述：以计算机可以处理的形式显示编码知识的问题。在专家系统时代，主要的方法是使用规则，尽管逻辑也被广泛应用于知识表述。

知识导航器：苹果公司在20世纪80年代推出的一款概念视频，引入了基于智能体的交互界面的思想。

知识工程师：受过构建知识系统训练的专业人士，知识工程师致力于研究如何进行知识获取。

知识获取：在建立专家系统时，从相关人类专家中提取和编码专家知识的过程。

知识库：在专家系统中，知识库由人类专家的知识组成，通常以规则的方式编码。

知识图谱：谷歌开发的一个庞大的基于知识的系统，自动从万维网上获取知识构建。

中文房间：哲学家约翰·希尔勒提出的一种设想，试图以此证明强人工智能不可能存在。

轴突：神经元与其他神经元连接的组成部分。另见突触。

侏儒问题：心理理论中的一个经典问题，当我们不经意间将心智的问题用另一个心智来解释的时候，就会出现这个问题。

智能体：一种功能比较完善的人工智能系统，通常集成了多种不同的人工智能功能，以便代表使用者自主工作。智能体通常被假定嵌入某个环境中，并在该环境中积极工作。

专家系统：一种使用人类专家知识来解决严格限制范围内问题的系统。典型的专家系统有MYCIN、DENDRAL和R1/XCON。从20世纪70年代末到80年代中期，构建专家系统是人工智能研究的重点。

自然语言理解：程序可以用人类的自然语言(如英语)进行交互。

组合爆炸：在连续选择中，选择的可能性等于选择的次数与备选可能性相乘。这是人工智能所面临的一个基本问题，在搜索中出现，它会

导致搜索树以极快的速度增长。

附录

附录A 理解规则

以下是动物例子的完整基础规则，我相信是由帕特里克·温斯顿(Patrick Winston)和伯索尔德·霍恩(Berthold Horn)提出的：

1. 如果动物有毛发，那么该动物是哺乳动物。
2. 如果动物能产奶，那么该动物是哺乳动物。
3. 如果动物有羽毛，那么该动物是鸟类。
4. 如果动物会飞，并且动物能产卵，那么该动物是鸟类。
5. 如果动物吃肉，那么该动物是食肉动物。
6. 如果动物是哺乳动物，并且动物有蹄，那么该动物是有蹄类动物。
7. 如果动物是哺乳动物，并且动物是食肉动物，并且动物毛发是黄褐色，并且动物有黑色条纹，那么该动物是老虎。
8. 如果动物是有蹄类动物，并且动物有黑色条纹，那么该动物是斑马。

让我们看看这些规则如何在正向推理下工作的。假设用户提供的信息是动物产奶，有蹄，有黑色条纹。然后正向推理将按如下方式进行：

1. 动物产奶这个事实会触发规则2，工作存储器里面会增加“该动物是哺乳动物”的信息。
2. 新增的动物是哺乳动物的信息，再加上动物有蹄，则触发规则6，工作存储器里面会增加“该动物是有蹄类动物”的信息。
3. 新增加的该动物是有蹄类动物的信息，再加上动物有黑色条纹的事实，则会触发规则9，工作存储器里会增加“该动物是斑马”的信息。
4. 到了这一步，没有进一步的规则可以触发了。

这样，我们就可以从动物的信息和规则中获得三项新的知识。

对于反向推理，我们从一些想假设的结论(目标)开始，再使用规则来回溯到更小的知识项。让我们看看，如果用户向系统呈现的目标是确定动物是否有蹄类动物，而没有给出任何信息，系统将如何工作：

1. 推理机试图寻找一个“该动物是有蹄类”的规则，在本例中，它找到了规则6。
2. 推理机没有足够的信息来触发规则6，因为它不知道前提(“动物是哺乳动物，动

物有蹄”)是否正确。因此，它给自己设立了一个目标，确认前提陈述是否属实。从技术上讲，“该动物是哺乳动物”和“该动物有蹄”成为子目标。

3. 首先来看“该动物是哺乳动物”这个子目标，在本例中，能够寻找到规则1和规则2。这就告诉推理机，这个目标可以通过两种方式达成：要么发现动物有毛发(规则1)，要么发现动物产奶(规则2)。因此，推理机取其中的第一个(“动物有毛发”)并将其作为子目标。

4. 推理机查找以“该动物有毛发”为结果的规则，然而，在本例中不存在这样的规则。我们已经到了规则回溯的尽头，但这并非死胡同，面对这样的情况，推理机可以向用户询问特定的语句是否为真。这样，使用反向推理过程就走到了向用户询问信息的阶段。现在我们假设用户并没有关于动物是否有毛发的信息，因此在回答推理机询问的时候会说“不知道”。在没有任何相关信息的情况下，推理机得出的结论是规则1不成立，因此它会继续以“该动物是哺乳动物”作为结论的规则2的前提进行推理，即以“动物能产奶”作为前提。

5. 推理机将“动物能产奶”作为子目标，然后发现这也是一个回溯到了尽头的目标。因此它同样会向用户提出询问，假设用户的回答为“是”，“动物能产奶”的信息会被添加到工作存储器，然后触发规则2，“该动物是哺乳动物”的信息被添加到工作存储器。

6. 到此，推理机已经建立好了规则6前提的第一部分，所以它继续第二部分，试图确定“动物有蹄”是否正确，因此，它将此作为子目标。

7.“动物有蹄”同样也是回溯尽头的规则，因此推理机会再次向用户询问。假设用户的回答为“是”，这一事实就会被添加进工作存储器，再加上“该动物是哺乳动物”的信息，系统就能触发规则6，最终得出结论，该动物是有蹄类动物。达成用户的原始目标。

附录B 理解PROLOG

假设我们想要一个解释家庭关系的系统，我们将了解一个简单的PROLOG程序如何捕获有关家庭关系的知识。假设我们用“female （ X ）”来表示 X 是女性，我们用“parents（ X, M, F ）”表示 X 的双亲为 M 和 F，其中 M 为父亲，F 为母亲。然后你就可以编写如下PROLOG规则：

sister_of（ X, Y ）: - female（ X ）, parents（ X, M, F ）, parents（ Y, M, F ）.

以上规则表述的是“X是Y的姐妹”，如果以下规则成立：

1. X是女性。
2. X的父亲是 M，母亲是 F。
3. Y的父亲是 M，母亲是 F。

如果你不熟悉逻辑推理，这听起来似乎是用相当复杂的方式在表述某人是某人的姐妹。但基本上，它表述的意思是，如果X是女性，X和Y有相同的父母，则X是Y的姐妹。

我们可以给PROLOG程序增加一些新的事实：

female (janine).

parents (janine, wayne, yvonne).

parents (david, wayne, yvonne).

鉴于这些事实，如果我们让PROLOG去证明Janine是David的姐妹，证明成立。本例中的相关目标是：

sister_of (david, janine)

当提出这个目标时，PROLOG就会回答“是”，表示它能够证明Janine是David的姐妹。

附录C 理解贝叶斯定理

我们在此继续讨论第四章中流感的例子，看看贝叶斯定理是如何起作用的。为了了解贝叶斯定律在这个例子中的应用，我们需要引入数学家称之为“符号”的知识。

首先，我们关注的假设是你罹患了流感，我们的证据是你的流感检测报告呈阳性。

现在，我们要计算的是先验概率，即你确实罹患流感的概率，计算根据是证据，即你的流感检测报告呈阳性。我们把假设成立的概率值写为Prob(假设|证据)。竖线“|”表示“基于”。

我们用Prob(假设)来表示假设(即你罹患流感)为真的先验概率。“先验概率”是指我们没有得到新的证据之前，认为假设成立的概率。在本例中，这就是随机挑选一个人，他得流感的概率。我们已知每1000人中大约有1个人感染流感，所以这个先验概率为1/1000 ＝ 0.001。我们可以把这个值看作假设成立的最初概率。如果我们对某人一无所知，那么对他罹患流感的概率估计就是随机挑选一个人，他罹患流感的概率就是千分之一。贝叶斯定理允许我们根据新的证据来更新这个结果，这就是它的魔力所在。

接下来，我们用Prob(证据|假设)来表示如果假设为真(即你罹患流感)，我们看到证据(即检测结果呈阳性)的概率。我们知道这个测试的准确率为99%，所以如果你罹患流感，那么在100次测试中有99次是呈现阳性的——如果你感染了流感，你的检测结果呈阳性的概率是0.99。

最后，我们用Prob(证据)来表示随机选择一个人，测试结果呈阳性的概率。计算出这个值是唯一非常棘手的部分。在本例中，这个概率值的计算方法是一个人罹患流感的概率乘以他的测试结果是阳性的概率，再加上他们没有流感的概率乘以测试仍然能给出阳性结果的概率。计算结果如下：

$$\text{Prob}(\text{证据}) = (0.001 \times 0.99) + (0.999 \times 0.01) = 0.11$$

那么，现在采用贝叶斯推断计算：

Prob(假设|证据)=[Prob(证据|假设)×Prob(假设)]/ Prob(证据)

代入我们刚才计算出的结果，即：

Prob(假设|证据)=[Prob(证据|假设)×Prob(假设)]/ Prob(证据)=(0.99×0.001)/ 0.011 = 0.09

所以，最终计算出来，你罹患流感的概率只有0.09，比十分之还稍低一点点。

附录D 理解神经网络

首先，我们来深入研究一下明斯基和帕普特强调的感知器模型的问题。单层感知器无法实现一个很简单的运算，被称为“异或”(XOR)。我们先假设一个有两个输入的单层感知器，如下所示：

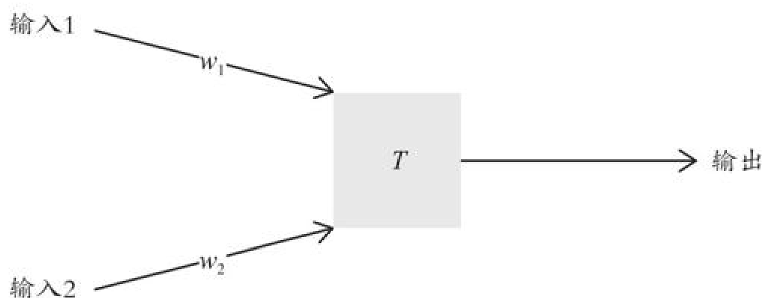


图22　单层感知器，无法计算异或

而异或函数是这样的：如果输入1激活，或者输入2激活，则触发感知器输出。而输入1和输入2同时激活或者同时不激活，都不触发输出。要理解为什么单层感知器无法计算异或函数，首先，我们假设它可以计算，那么，对于输入1的权重值w_1，输入2的权重值w_2，以及感知器的触发阈值T，它们必须满足如下要求：

·如果两个输入均未被激活，则神经元接收到的输入权重为0，此时神经元要保持不被触发，因此触发阈值T必须大于0。

·如果只有输入1被激活，则感知器被触发，因此w_1必须大于触发阈值T。

·如果只有输入2被激活，则感知器被触发，因此w_2必须大于触发阈值T。

·如果输入1和输入2都被激活，则感知器不会被触发，因此$w_1 + w_2$的值必须小于触发阈值T。

但很容易就能看出，满足上述所有要求的w1、w2和T值不存在，因此，感知器无法计算异或函数。

现在让我们看看两层感知器结构如何计算异或函数，如图23所示。

我可以肯定地说，这个感知器网络能够正确计算异或函数。它由三个神经元组成：当输入1或者输入2被激活时，神经元1会被触发；当输入1和输入2没有同时被激活，神经元2才会被触发；最后，如果神经元1和神经元2都被触发，神经元3才会被触发。总而言之，输入1和输入2有一个被激活时，神经元3被触发；而输入1和输入2同时被激活或者同时未被激活，神经元3都不会被触发。

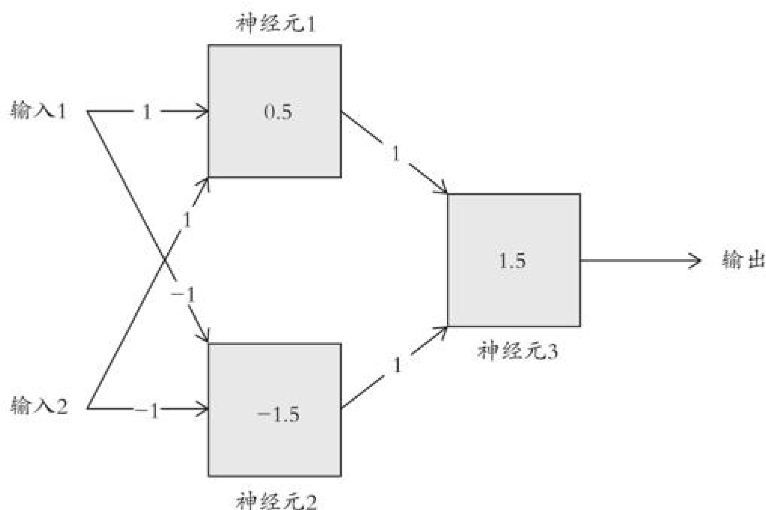


图23　两层感知器结构可以计算异或函数

更详细地说，就是考虑如下四种情况：

1. 假设输入1和输入2都未激活，此时神经元2将被触发(因为它接收的输入O高于-1.5的触发阈值)，但神经元1不会被触发。这种情况下，神经元3接收到的输入为1，低于触发阈值1.5，因此神经元3不会被触发。

2. 假设输入1激活，输入2未激活。神经元1接收到的输入权重为1，高于触发阈值0.5，则神经元1被激活，神经元2接收到的输入权重为-1，高于触发阈值-1.5，同样，神经元2也被触发。此时神经元3接收到神经元1和神经元2的输入，权重值为2，高于触发阈值1.5。因此神经元3会被触发。

3. 假设输入2激活，输入1未激活。与情况2相同的计算方式计算得出，神经元3会被触发。

4. 假设输入1和输入2同时激活，此时神经元1接收到的输入权重为2，高于触发阈值0.5，则神经元1被激活。神经元2接收到的输入权重为-2，低于触发阈值-1.5，则神经元2不会被触发。神经元3接收到的输入权重为1，低于触发阈值1.5，因此神经元3不会被触发。

延伸阅读

或许有些读者希望通过延伸阅读来满足他们的好奇心，下面列举的是一些关键的文献。

现代学术界对人工智能的权威介绍请参见斯图亚特·罗素和彼得·诺维格合著的《人工智能：现代方法》(第三版，培生出版集团，2016年)。这是我对人工智能技术的标准参考书，我在撰写本书的时候反复使用它来核对技术细节。

如果你对历史细节和人工智能领域的发展方式感兴趣，推荐你读尼尔斯·尼尔森的《人工智能探索》(剑桥大学出版社，2010年)。这是一本极佳的现代人工智能发展历史指南，由该领域最伟大的研究人员之一撰写。

关于艾伦·图灵生平的书有不少，但迄今为止最好的一本讲述图灵生平的书是安德鲁·霍奇斯所著的《艾伦·图灵传：如谜的解谜者》(贝内特图书/哈钦森出版社，1983年)。我读大学的时候非常喜欢威廉·考夫曼和豪里斯工程出版社出版的三卷本《人工智能手册》(卷一，艾夫伦·巴尔和爱德华·费根鲍姆编辑，1981年；卷二，巴尔和费根鲍姆编辑，1982年；卷三，保罗·科恩和爱德华·费根鲍姆编辑，1982年)。它们描绘了黄金年代建造各种系统及所开发技术的光辉画面，虽然绝版好一阵了，但是网上也能买到便宜的复印件。彼得·杰克逊的《专家系统导论》(艾迪森·威斯利出版公司，1986年)是研究专家系统极好的参考文献，这是30年前我上学时候的必读书，很高兴在撰写本书的时候重温它。有关专家系统知识工程的详细介绍，请参阅《构建专家系统》(弗雷德里克·海耶斯-罗斯、唐纳德·沃特曼和道格拉斯·莱纳特编辑，艾迪森·威斯利出版公司，1983年)。要想理解人工智能中的逻辑传统，我推荐迈克尔·杰纳西雷斯和尼尔·尼森的《人工智能逻辑基础》(摩根·考夫曼出版社，1987年)。我在研究生时期仔细阅读过本书，虽然它所支持的技术在最近十年内基本上过时了，但它仍然是人工智能研究领域最有影响力的著作之一，也是介绍逻辑学的优秀著作之一。

1999年，罗德尼·布鲁克斯在著作《寒武纪智能》(麻省理工学院出版社，1999年)中收集了他的论文。我厚着脸皮推荐自己的书《多智能

体系统引论》(第二版，威利出版公司，2009年)，作为对智能体、智能体结构和多智能体系统的介绍。要了解更多关于博弈论在人工智能中的应用，请阅读《多智能体系统：算法、博弈论和逻辑基础》，作者是约夫·肖姆和凯文·莱顿-布朗(剑桥大学出版社，2008年)。

机器学习是一个很浩瀚的话题，我在正文里提及的仅仅是皮毛。虽然我把重点放在神经网络上——因为它是目前最主流的研究领域，但还有不少其他的机器学习技术。如果你想了解更详细但相对非正式的介绍，请参阅埃塞姆·阿培丁的《机器学习：新人工智能》(麻省理工学院出版社，基础知识系列，2016年)。如果你想在深度思维找一份工作(这年头，谁不想呢)，那么你可能需要读一下《深度学习》(麻省理工学院出版社，2017年)，作者是伊恩·古德费洛、约书亚·本吉奥和亚纶·库维尔。最后，请注意，罗素和诺维格的著作(我在第二段提到过)也提到过不少机器学习领域的好方法。

关于奇点的讨论，我推荐默里·沙纳汗所著的《技术奇点》(麻省理工学院出版社，基础知识系列，2015年)；关于人工智能伦理学，请参阅维吉尼亚·迪格努姆的《负责任的人工智能》(施普林格出版社，2019年)；关于技术和就业的问题，请参阅卡尔·贝内迪克特·弗雷的《技术陷阱》(普林斯顿大学出版社，2019年)。

玛吉·博登的《人工智能与自然人》(麻省理工学院出版社，1977年)是一本介绍强人工智能和有意识机器的经典著作。有关意识的详解教科书，请参阅苏珊·布莱克莫尔的《意识概论》(霍德&斯托顿出版公司，2010年)。丹尼尔·丹尼特写的关于思维、意识和人工智能的文章几乎都值得一读，我推荐《心智类型》(基本书局，1996年)作为入门，这本书之后，推荐你再读一篇文章《我在哪儿》，最初出现在丹尼特1978年出版的《头脑风暴》一书中，这本书在网上很容易找到。

作者注

许多关于人工智能和机器学习的当代文章都存放在一个名为

“arXiv”的开放在线文档库中。arXiv中的文章使用简单的编号方案，类似于arXiv：1412.6572。我们可以通过访问https://arxiv.org并输入编号(在本例中为“1412.6572”)来获取相关文章。

序言

1. http://tinyurl.com/y7zc94od

2. http://tinyurl.com/yxk3xurl

第一章　图灵的电子大脑

[1]. 霍奇斯，《艾伦·图灵传：如谜的解谜者》，贝内特图书/哈钦森出版社，1983年。

[2]. 除了惊人的科学成就，图灵在英国还有着深远的社会影响。经过长期公开高调的运动，英国政府在2014年对他进行了赦免。不久之后，所有根据同一法律被起诉的人都获得了赦免。

[3]. 这是检查质数最直观的方法，但绝不是最优雅或最有效的方法。自古以来，人们就知道埃拉托斯特尼筛法（由希腊数学家埃拉托斯特尼提出的一种简单检定质数的算法。）更为简单清晰。

[4]. 此后我将不再区分图灵机和通用图灵机，都用图灵机来表述。

[5]. 图灵将解决判定问题的荣耀与普林斯顿大学数学家阿隆佐·邱奇分享，后者先于图灵独立获得了一个截然不同的结果证明。然而，图灵的证明被认为是决定性的：它更直接、更完整、更易懂，而且影响巨大。他据此发明了图灵机，改变了世界。

[6]. 严格地说，算法就是一种方法，而程序是一种用实际编程语言(比如Python或者Java)编码的算法。因此，算法独立于编程语言。

[7]. 图灵机的编程实际上更为原始，我在这里列出的指令是典型的相对低级的编程语言，但仍然比图灵机程序所使用的抽象得多(也更容易理解)。

[8]. 科尔曼和李维斯特，《算法导论》(第一版)，麻省理工学院和麦格劳-希尔出版社，1990年。

[9]. 图灵，《计算机器与智能》，《心智》，40，1950年，第433-460页。

[10]. 这段对话是由苹果麦金塔电脑附带的ELIZA版本生成的。如果你有苹果电脑，可以自己试试。打开Applications文件夹，再打开里面的Uti lties文件夹，然后双击终端应用程序图标来启动终端程序。在终端窗口会出现一大堆傻乎乎的东西，当它稳定下来以后，按Esc键(在键盘左上角)，然后按X键，然后键入“doctor”，再按回车键。好啦，看看吧。但请记住：它不是真的!

[11]. http://tinyurl.com/y7nbo58p

[12]. 不幸的是，文献中的术语并不精确，也不统一。大多数人似乎用“通用人工智能”指代能够产生类似人类智能行为的机器，而不关心诸如它们是否具有自我意识之类的哲学问题。从这个意义上说，通用人工智能大致相当于希尔勒所指的弱人工智能。然而，混淆视听的是，有时候这个词又被用来表示更像希尔勒所指的强人工智能。在本书中，我用它来表示弱人工智能。

[13]. http://tinyurl.com/y76xdfd9

第二章　黄金年代

[14]. 其中最有影响力的技术来自约翰·麦卡锡的一种叫作分时的概念。他意识到，人们使用计算机的时候，大部分时间计算机都是空闲的，等待人们输入东西或者运行程序。他意识到这个“空闲时间”可以和其他用户共享，允许多人同时使用计算机。这个概念使昂贵的计算机得到更有效的利用。

[15]. 实际上它代表“Lisp处理器”。LISP进行符号运算，而符号列表正是实现这一点的关键。

[16]. 娜萨，《美丽心灵》，西蒙与舒斯特出版公司，1998年。

[17]. 麦卡锡等人，《关于达特茅斯夏季人工智能研究项目的建议》，1955年(转载于《人工智能》，24(4)，2006年，第12-14页)。

[18]. 威诺格拉德，《理解自然语言》，学术出版社，1972年。

[19]. 在语言学中，用“他”“她”和“它”等词来指代先前在对话中出现的实体，称为复指。一个试图理解或建立自然语言对话的计算机程序必须能够解决复指指代的具体对象问题，这在当今仍然是一个挑战。SHRDLU(有限的)处理回指能力被认为是突破性的。

[20]. 菲克斯和尼森，《STRIPS：定理证明应用于问题求解的新途径》，《人工智能》，2(3-4)，1971年，第189-208页。

[21]. 控制SHAKEY的电脑是一台PDP-10，20世纪60年代末最先进的主机电脑，它重达1吨多，需要一个大房间才能容纳。一台PDP-10可以配备高达1兆字节的内存——我口袋里的智能手机内存容量比它高大概4000多倍，而且速度快得不可思议。

[22]. http://tinyurl.com/yxu8hwoq

[23]. http://tinyurl.com/n6lf8t6

[24]. 这并非精确的数字，只是一个近似值，让我们对涉及的数字比例有一些了解。

[25]. 从技术上讲，假设b为搜索问题的分支因子，d是搜索树的深度，如此，搜索树的底层(深度为d的那一层)将包含bd种状态，即b的d次幂，b×b×b×b×……×b(d次)。分支因子表现出的增长速度在技术上称为指数型增长，一些参考文献使用几何级数增长这一术语，尽管我认为我从来没有在人工智能领域相关文献见过这个术语。

[26]. 跳棋游戏，美国称之为“checkers”，英国称之为“draughts”，因为塞缪尔是美国人，所以作者表述跳棋的原文用了“checkers”。另外，在人工智能领域，这个程序通常被称为“塞缪尔的跳棋程序”(Samuel's checkers player)，如果把它称为“Samuel's draughts player”会显得很怪异。

[27]. 哈特、尼尔森和拉斐尔，《最小成本路径启发式测定的正式基础》，《IEEE系统科学与控制论汇刊》，4(2)，1968年，第100-107页。

[28]. 类似这样的计算问题都被命名了，本例叫作“独立集”。

[29]. 正文中旅行推销员问题是简化描述，更精准的描述如下：有一个城市列表C，对于C之中的每一对城市i和j，我们都有一个距离di, j，定义di, j等于i和j之间的距离。另外，存在一个“上限”B，这是旅行商在消耗完燃料之前可以行驶的总路程。现在我们要回答的问题是，是否有一个游览所有城市的方案(即用某种方式对C中的元素进行排序)，使得按照这个方案从每一个城市到下一个城市，最终回到出发时的城市，总行程不超过B。

[30]. P代表多项式时间，在多项式时间内能够运行完成的算法，才是解决问题的可行算法。如果你想了解更多有关NP完全问题和P与NP问题的知识，可以参见延伸阅读中提到的参考文献。

第三章　知识就是力量

[31]. 温斯顿和霍恩，《LISP》(第三版)，培生出版集团，1989年。

[32]. 肖特利夫，《基于计算机的医学诊断系统：MYCIN》，美国爱思唯尔出版集团，1976年。

[33]. 尚克和亚伯森，《脚本、计划、目标和理解：对人类知识结构的探究》，心理学出版社，1977年。

[34]. 伍兹，《链接中有什么？语义网络基础》，博罗和A·柯林斯编，《认知科学中的表征和理解研究》，摩根·考夫曼出版社，1975年。

[35]. 麦克德莫特，《塔斯基语义学，或称无意指不表达!》，《认知科学》，

2(3)：第277-282页，1978年。这个标题有点像美国独立战争时期的口号“无代表不纳税”(麦克德莫特就是在1976年美国独立200周年庆祝活动前后写了这篇文章)。

[36]. 麦卡锡，《逻辑人工智能的概念》，未发表。

[37]. 克罗克森和梅利什，《PROLOG编程》，施普林格出版社，1981年。

[38]. PROLOG中使用的演绎形式被称为解析，最初是在20世纪60年代出现的，它可以有效实现PROLOG中使用的规则。

[39]. 沃伦，《生成有条件的计划和编程》，《第二届夏季人工智能行为与模拟会议论文集》(AISB-76)，爱丁堡，1976年7月。

[40]. 古哈和莱纳特，《Cyc：中期报告》，《人工智能杂志》，11(3)， 1990年。

[41]. 普拉特，《CYC报告》，未发表，1994年，访问地址：http://tinyurl.com/y4q4aoqj

[42]. 赖特，《默认推理逻辑》，《人工智能》， 13， 1980年，第81-132页。

[43]. 这个例子被称为尼克松菱形，用图形表示的话，图形呈菱形。

第四章　机器人与其合理性

[44]. 布鲁克斯，《无表征智能》，《人工智能》， 47， 1991年，第139-159页。

[45]. 有趣的是，人类智力的这些更高级的方面是由大脑内叫作新皮质的部分来处理的。而人类进化的记录告诉我们，大脑新皮质部分的进化相对时间靠后。对于人类来说，推理和解决问题是比较新兴的能力：在进化史的大部分时间里，我们的祖先都不具备这些能力。

[46]. 罗素和苏布拉曼尼亚，《可证明的有边界最优智能体》，《人工智能研究杂志》， 2， 1995年。

[65]. https://www. irobot.co.uk

[47]. 布鲁克斯，《一个走路的机器人：进化网络的应急行为》，《机器人与自动化会议论文集》， 1989年，亚利桑那州斯科茨戴尔，1989年5月。

[48]. 弗格森，《旅行机：有态度的自主智能体》，《IEEE计算机》，25(5)，第51-55页，1992年。“旅行机”这个名字显然是在隐喻

“图灵机”〔英文中“Turing Machines”(图灵机)和“Touring Machines”(旅行机)读音一样〕。25年来，旅行机的开发者因内斯·弗格森一直是我的好朋友，但我不确定是否要原谅他。如果他知道30年后的

人们还在写这种笑话，或许他就不会觉得这么好笑了。

[49]. 维尔和比克莫尔，《基础智能体》，《计算智能》，6，1990年，第41-60页。

[50]. 出于历史准确性的考虑，我应该指出，图形用户界面和桌面界面的雏形并不是由苹果公司发明的。这应该归功于施乐公司帕洛阿尔托研究中心(PARC)的研究人员。然而，苹果公司认识到其潜力，并将其制成了产品。

[51]. http://tinyurl.com/y9qxdko5

[52]. 梅斯，《智能体能够帮助人们减少工作量和信息过载》，《ACM通讯》，37(7)，1994年，第30-40页。

[53]. 埃齐奥尼和威尔德，《基于软件机器人的互联网交互》，《ACM通讯》，37(7)，1994年，第72-76页。

[54]. 冯·诺依曼和摩根斯坦，《博弈论与经济行为》，普林斯顿大学出版社，1944年。

[55]. 为简单起见，我把钱和效用等价。在实践中，货币和效用也是经常相关的，但它们并不是同一概念，如果你认为经济效用只是与钱有关，那么你会惹恼诸多经济学家。实际上，效用理论只是用数值的方式捕捉和计算偏好。

[56]. 罗素和诺维格，《人工智能：现代方法》(第三版)，培生出版集团，2016年，第611页。

[57]. 墨菲，《人工智能机器人简介》，麻省理工学院出版社，2001年。

[58]. 珀尔，《智能系统中的概率推理：合理推理网络》，摩根·考夫曼出版社，1988年。

[59]. 伍尔德里奇，《多智能体系统引论》(第二版)，威利出版公司，2009年。

[60]. 鲁宾斯坦和奥斯本，《博弈论课程》，麻省理工学院出版社，1994年。

[61]. 塞尔曼、莱维斯克和米契尔，《一种解决难满足性问题的新方法》，第十届国际人工智能会议论文集(AAAI_1992)，美国加利福尼亚州圣何塞，1992年。

第五章　深度突破

[62]. 深度思维的收购价各个媒体的报道不尽相同，《卫报》报道的数字为4亿英镑。http://tinyurl.com/kvyueye

[63]. 明斯基和帕普特，《感知器：计算几何导论》，麻省理工学院出版社，1969年。

[64]. http://tinyurl.com/ycu4ngsg

[81]. 鲁梅尔哈特和麦克莱兰(编)，《并行分布式处理》(2卷)，麻省理工学院出版社，1986年。

[66]. 鲁梅尔哈特、辛顿和威廉姆斯，《利用反向传播错误学习表征》，《自然》，323， 1986年，第533-536页。

[67]. 古德费洛，本吉奥和库维尔，《深度学习》，麻省理工学院出版社，2016年。

[68]. 从历史增长率来看，我们可以预期在40年左右的时间里，神经网络能达到与人脑相同数量的人工神经元。不过，这并不意味着人工神经网络将在40年内实现人类水平的智能，因为大脑不仅仅是一个神经网络，它还有结构。

[69]. http://www.image-net.org

[70]. https://wordnet.princeton.edu

[71]. 克里泽夫斯基、苏茨科弗和辛顿，《基于深度卷积神经网络进行网络图像分类》，神经信息系统大会，2012年，第1106-1114页。

[72]. 古德费洛等人，《解释和利用对抗性案例》， arXiv: 1412.6572。

[73]. 姆尼赫等人，《使用强化深度学习来玩雅达利游戏》， arXiv: 1312.5602v1。

[74]. 姆尼赫等人，《通过强化深度学习实现人类水平控制》，《自然》，518，2015年，第529-533页。

[75]. 西尔弗等人，《使用深度神经网络和搜索树掌握围棋游戏》，《自然》，529， 2016年，第484-489页。

[76]. http://tinyurl.com/ydafuhj p

[77]. 西尔弗等人，《脱离人类知识干预掌握围棋游戏》，《自然》，50， 2017年，第354-359页。

[78]. https://www.captionbot.ai

[79]. https://translate.google.com

[80]. 这一段是由苏格兰作家兼翻译家斯科特·蒙克利夫翻译的，他的翻译是文学史上最著名的译本，被认为是不逊于原著的杰作。尽管也有人批评他滥用了普鲁斯特的语言。

第六章　人工智能的今天

[82]. https://tinyurl.com/y2k5aeq4

[83]. https://tinyurl.com/y8bu8xx8

[84]. https://tinyurl.com/y5y75rgs

[85]. https://blog.cardiogr.am/tagged/research

[86]. 这个数字可能会受到生产过程中死亡以及夭折的人口比例影响，一位成年人有机会活到我们现在所认为合理的年龄。

[87]. http://tinyurl.com/yc5gv8jg

[88]. 德法乌等人，《深度学习在视网膜疾病诊断及转诊方面的临床应用》，《自然医学》，24，2018年，第1342-1350页。

[89]. http://tinyurl.com/yakkuyg2

[90]. 赫尔曼、布伦纳和斯塔德勒，《自动驾驶》，爱莫瑞德出版社，2018年。

[91]. https://corporate.ford.com/innovation/autonomous-2021.html

[92]. https://www.riotinto.com/media/media-releases-237_23991.aspx

第七章　杞人忧天——我们想象中的人工智能会出什么错

[93]. http://ti nyurl.com/ybsrkr4a

[94]. 库兹韦尔，《奇点临近》，企鹅出版集团，2005年。

[95]. 温格，《即将到来的技术奇点：如何在后人类世代生存》，美国宇航局刘易斯研究中心，《21世纪展望：网络空间世代的跨学科科学与工程》，第11-22页。

[96]. 沃尔什，《奇点可能永远不会临近》，arXiv:1602.06462v1。

[97]. https://tinyurl.com/y622vm6k

[98]. 威尔德和埃齐奥尼，《机器人学第一定律：召唤武器》，《国际人工智能会议论文集》，1994年，第1042-1047页。

[99]. 富特，《堕胎问题和双重效应理论》，《牛津评论》，第5期，1967年。

[100]. http://tinyurl.com/ybl8l uoe

[101]. 阿瓦德等人，《道德实验机器》，《自然》，563，2018年，第59-64页。

[102]. http://tinyurl.com/ydf26689

[103]. http://ti nyurl.com/jslm95f

[104]. 为了透明起见，我应该说明，两次促成阿西洛玛人工智能准则的会议都曾邀请我参加，我本来很想去。可惜很不凑巧，两次都因为承诺了别的事情而不得不食言。

[105]. http://tinyurl.com/y28osmtw

[106]. http://tinyurl.com/y29v4rrd

[107]. http://tinyurl.com/yc3vgkgv

[108]. http://tinyurl.com/y2egvzxx

[109]. 迪格努姆，《负责任的人工智能》，施普林格出版社，2019年。

[110]. 博斯特罗姆，《超级智能》，牛津大学出版社，2014年。

[111]. 汉森，《什么是平等偏好》，《哲学逻辑杂志》，25(3)，1996年，第307-332页。

[112]. 吴恩达和罗素，《逆向强化学习算法》，《第十七届机器学习国际会议论文集》，2000年。

第八章　现实中的人工智能会导致什么问题

[113]. 贝内迪克特·弗雷和奥斯本，《就业的未来：电脑化将如何影响工作》，《技术预测与社会变革》，114，2017年1月。

[114]. https://rodneybrooks.com/blog/

[115]. https://tinyurl.com/yytefewg

[116]. http://tinyurl.com/ydb9bpz4

[117]. http://tinyurl.com/ycq6jk35

[118]. http://tinyurl.com/y74yfk8a

[119]. 奥斯瓦尔德等人，《算法风险评估-治安模型：从杜伦HART模型与“实验性”相称中所学到的经验》，《信息与通信技术法》，27：2，2018年，第223-250页。

[120]. http://tinyurl.com/y6narok3

[121]. https://www.predpol.com

[122]. http://tinyurl.com/y242nn5u

[123]. http://tinyurl.com/ycef9mqv

[124]. http://tinyurl.com/y4elgklp

[125]. 我认为这种场景是由著名人工智能专家斯图尔特·罗素提出的。

[126]. http://tinyurl.com/yy7szdxm

[127]. 阿尔金，《控制致命行为：将伦理嵌入混合审议/反应机器人架构中》，技术报告GIT- GVU- 07-11，佐治亚理工学院计算机学院。

[128]. https://www.stopki l lerrobots.org/

[129]. 上议院人工智能特别委员会，2017—2019届报告，《人工智能在英国：准备好了吗？愿意接受吗？可行吗？》，HL Paper 100，2018年4月。

[160]. http://tinyurl.com/l btnkse

[130]. https://tinyurl.com/y9juww8v

[131]. https://tinyurl.com/y7dzz46v

[132].《自然》，563，2018年11月27日，第610-611页。

[133]. 卡罗琳·克里亚多·佩雷斯，《看不见的女人：在为男性设计的数据里暴露偏见》，查托温德斯出版社，2019年。

[134]. http://tinyurl.com/y9cd9x7f

[135]. http://tinyurl.com/y25dhf9k

[136]. 脸书允许你查看他们为你建立的偏好图。http://tinyurl.com/j4ys4hq

[137]. http://tinyurl.com/y7mcrysq

[138]. https://tinyurl.com/yyc6botm

[139]. http://tinyurl.com/yaypy567

[140]. http://tinyurl.com/yd36fdva

[141]. http://tinyurl.com/y6uoewyg

[142]. http://tinyurl.com/y8vgslkb

[143]. http://tinyurl.com/y6wx5tz7

第九章　通往有意识的机器之路

[144]. http://ti nyurl.com/yxwlrrkq

[145]. 内格尔，《成为蝙蝠是什么样的感受》，《哲学评论》，83：4，1974年，第435-450页。

[146]. 卡尼曼，《思考：快与慢》，企鹅出版集团，2012年。

[147].“谁能理解车夫，并能控制自己的思想，他就会到达旅程的终点，那是无所不在的至高无上的居所。”(《卡达奥义书》，1.3)

[148]. 苏恩等人，《人脑自由决策的无意识因素》，《自然：神经科学》，11，2008年，第543-545页。

[149]. 更确切地说，邓巴感兴趣的是大脑皮层的大小。新皮层是大脑处理感知、推理和语言的部分。

[150]. 丹尼特，《意向立场》，麻省理工学院出版社，1987年。

[151]. 丹尼特，《认知行为学中的意向系统》，《行为与脑科学》，6， 1983年，第342-390页。

[152]. 肖汉，《面向智能体编程》，《人工智能》，60(1)， 1993年，第51-92页。

[153]. 麦卡锡，《机器具有心理因素》， V·利夫希茨(编辑)，《形式化常识：约翰·麦卡锡论文集》，阿尔布利克斯出版社，1990年。

[154]. http://ti nyurl.com/yc2knerv

[155]. 这里的关键词是“有意义”。每当有人提出类似测试时，总有人会试图找到某种方法来在测试中玩点诡计，这样他们就可以宣称自己成功了，即使是以一种你没有预料到的方式做到的。当然，这正是图灵测试所发生的事情。我所追求的是能够以实质性的方式实现这一点的程序，而不是利用烟幕弹和镜像回答的组合来通过测试的程序。

[156]. 西蒙·拜伦-科恩，A. M. 莱斯利和U. 弗里斯，《自闭症孩子拥有心智理论吗》，《认知》，21(1)， 1985年，第37-46页。

[157]. 西蒙·拜伦-科恩，《心智盲症：一篇关于自闭症和心智理论的文章》，麻省理工学院出版社，1995年。

[158]. 拉比诺维茨等人，《心智机器理论》， arXiv: 1802.07740。

[159]. 我不是进化心理学方面的专家：本节我的指南是罗宾·邓巴的《人类进化》(企鹅出版集团，2014年)，我很高兴向感兴趣的读者介绍更多细节。